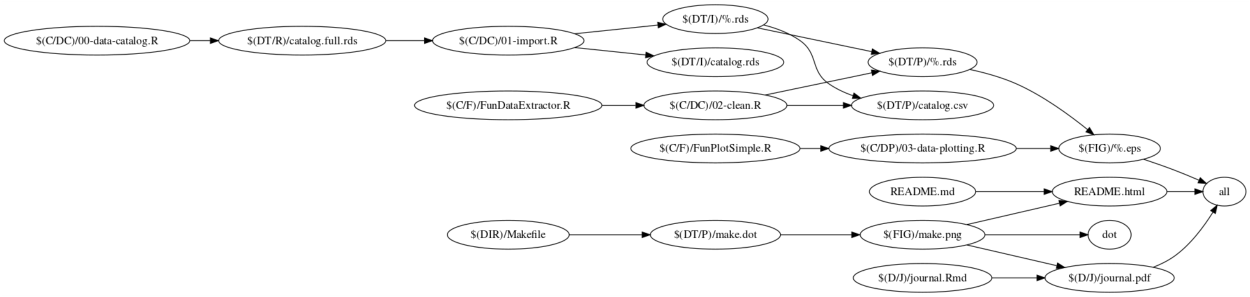
Journal

Tuesday 5.12.2017

1. This project had been started over a month ago, and got pretty far, then was asked by SH to stop working on the FS.. Anyway, I'm now starting again, first by migrating it into my new project template, makefile and all.

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



2. Initialise readme. And make it have the makefile plot. And the building of the readme be in the makefile plot. So meta ;)

3. Initial commit.

4. Migration:

- DHS application form -> dictionaries

5. Now think through how to deal with the data import, because there is like 2GB of this stuff zipped, let alone unzipped. OK, there seems to be a library that helps with this, called `lodown` and this is part of the **Analyse survey data for free project** by Anthony Joseph Damico [http://asdfree.com/prerequisites.html]. So let's try that first.

6. Requires new Rtools, then devtools package and finally `install_github( "ajdamico/lodown" , dependencies = TRUE )` to install the lodown package from github.

7. OK, that didn't work :(Got " length(project.line) == 1 is not TRUE" error when tried to run `get_catalog()`, no idea why.

8. OK, second try. So I start with a form on the DHS website, where i pick the countries, (all), the survey (individual recode) and the file type (stata system file). Then I get a url list. It seems that if I am logged in and in the correct 'project' in a browser, I can download from that browser. So presumably I won't be able to download fro R direct using `download.file()`.

9. Right, what I'll do is simply download all the files into data/raw, period. Here is the process:

- Get the text file with the urls from the DHS website. saved in `data\raw\url.list.txt`.
- download all the (currently) 321 zip files into `data\raw`. Shit, I've got 318 of them already, which ones are missing?
- WOW, so turns out that the 'unique file naming system' is actually not unique at all, since the Indian surveys are actually split into states and (at least) one of these has the same name as Kenya `KEIR42DT.zip`. FFS. OK, so now I have to rename the Kenya ones as KN. MANUALLY.
- Then prepare a dataframe - catalog - to deal with the importers etc.
- learn regular expressions to get out the survey ID, which helps distinguish the Indian ones.

DHS DOCUMENTATION LINK

Wednesday 6.12.2017

1. OK, continuing with the `00-import.R` file, which should output individual `.rds` files, and a table summarising the datasets. Input is the url list (and the files already downloaded in the `/data/raw` folder.
2. This is the annoying manual shit: I have to fix the KE file extensions to KN, in the folder, and in the catalog dataframe.
3. During importation there are numerous warning messages for duplicate labels. We are summarily ignoring these.
4. Preparing a function to import and clean up after itself.. A bit complicated ;) The `Stata.file()` function - the importer - unzips the file into the working dir, i need to get rid of the files.. Also the .dta files have different names, i.e. cases..
5. Let it run over night.

Thursday 7.12.2017

1. Stopped over night because there was not enough disk space.. I couldn't automatically delete the .dta files, because the importer is some sort of external pointer that is kept open randomly and makes it impossible to delete the file, even though you then can a while later or sth.. Anyway.
2. Reran the rest of the files. It did get stuck again a little bit with some dta error - version 0 isn't supported, but that cleared up magically after I reran teh third time.
3. Looks like the last four or so importers keep an open connection.
4. Error in get.dictionary.dta(dta) : version 80 not yet supported, happened again.

Tuesday 12.12.17

1. Got word from the `lodown` package developer that I might want to look at the project name I have on the DHS website, and it looks like the length of the name made lodown not work. So actually if I change it to sth shorter it would work. [[https://github.com/ajdamico/lodown/issues/124]]
2. This is very annoying mainly because I've already done the whole thing manually in the meantime. Classic.

Thursay 14.12.17

1. Argh, for complete reproducibility let's go and do the `lodown` thing then.
2. What Anthony, the `lodown` guy might find useful is a regex thing to make he catalogue a bit more useful, because the catalog doesn't distinguish different types of files (surveys and file types).
3. Also, the lodown catalog for DHS puts every file in it's own folder it seems, so let's first try to make it not do that.
4. Right, seems that I extracted the dataset type and file types correctly, as i end up with 321 files in my catalog, same as I had in the manually ticked url list from the DHS website.
5. Now change the download folder. Test with single dataset.

6. OK, testing looks good! It's downloading the file, and unzipping it, and looks like it's also saving the .dta file as an .rds object.

Friday 15.12.17

1. Now need to get the `00-import.R` to source properly from the makefile.
2. So how can I stop the `00-import.R` script from importing the catalog every time? Ah, nice, separate the catalog build from the import. Now won't have to repeat it every time.
3. Now I have a different problem. The lodown package is great in extracting the file, but it reads the `dta` file and saves it as a dataframe, thereby loosing the value label data.
4. AAh, but it isn't, it uses `haven::read_dta`, so the labels are there, if only in some weird `labelling` format.. But OK, you have to use `haven::as_facto()` to get it to behave as a factor, all good.
5. Now, let's try and run the whole download thing, that should be the last of 01-import.R And let me commit this now as working download..
6. Started it, went to 80/226, then got a memory error… I've upped the memory limit now, hope that's all it needs. It's India, isn't it. It is, and it made it!! Hooray, let's not run this again ;)
7. In the mean time I write a new cleaning file. `/code/data-cleaning/02-clean.R`
8. Man, 185 of 226, I know what it is as well, the double zip file… https://userforum.dhsprogram.com/index.php?t=msg&goto=13740&#msg_13740 is an open issue, but ffd, I'm afraid I'll have to do this manually. From https://dhsprogram.com/customcf/legacy/data/download_dataset.cfm?Filename=SNIR7QDT.ZIP&Tp=1&Ctry_Code=SN&surv_id=524 And now the `01-import.R` file starts from 187 onwards, that's just the way it is…
9. OK, there's another Senegal like that as well, SNIR70DT.. Hope the rest are not fucked!
10. So I should end up with the zip files deleted from `/data/raw` so only two of the .dta Senegal files should be kept there in case this ever needs to be repeated.
11. The there should be 226 .rds interim files, and the one catalog.rds file there as well. These catalog.rds files are a bit annoying because I use the wildcard thing in the makefile to look at all of them - and see if they need updating, but since both `01-import` and `02-clean` produce both all the rds files at the same time, this is OK.
12. BUT it won't be OK if I put any other rds files in there. Technically I shouldn't, because I have the `/results` folder, so don't let me forget that!
13. Ach, and there is an error with fucking Kerala, since now I managed to write both KE and IK on the same file, but because the actual Kerala doesn't even have v632, I'll just pretend It didn't happen. (ToDo) ** file naming conventions DHS ** are here: https://dhsprogram.com/data/File-Types-and-Names.cfm
14. OMG, looks like `02-clean` also works -> never ever touch those two files again ;)
15. (NB: I'm not sure why the makefile wouldn't work until I added the second catalog to the all dependency..
16. Also, delete the "IK" one, cause it isn't Kerala anyway, just a double Kenya.
17. So with one catalogue and one make.dot, this leaves 144 .rds files in the processed folder.

Wednesday 20.12.17

1. Running on Linux now, so far so good. `make dot` works ;)

2. DHS forum question asked here: https://userforum.dhsprogram.com/index.php?t=rview&goto=13741#msg_13741
3. Opened Senegal Issue on github, Anthony asked for debug.
4. This means installing on this machine R 3.4, devtools, lodown.., but getting there.
5. And fixing the Senegal issue in the `lodown` package, how cool ;) [https://github.com/ajdamico/lodown/issues/126]
6. OK, now I need to get on the plotting! So we have 144 processed files, each whit 4 variables. Weights, rural/urban,
7. I'll just check the catalogue and v632 matches the files in `data\processed` to be sure.
8. OK, Senegal `lowdown` PR made, let's see ;) Learned a neat looping trick, where you remove the current object from the looping index

Thursday 21.12.17

1. Fix public credentials issue by using `.config` file in .gitignore
2. Let's see if I can remove the git traces of it as well! save 01-importation as new file, delete old one. Commit just the delete first. then `add -u` and then commit the new one.
3. That didn't work, I;ll try the [https://rtyley.github.io/bfg-repo-cleaner/] later.
4. OK, get on to plotting. First of all, it seems that the `02-clean` code lost the labelling that came along with the first rds. Is there an easy fix? Or do I ignore it for now?
5. Seems OK, from `labelled` package documentation: `Some base R functions like subset drop variable and value labels attached to a variable. copy_labels could be used to restore these attributes.`
6. Actually `subset` was not dropping labels, or rather it only did so because I changed the variable names. Changing them later fixes that.
7. Rerun the `02-clean.R` on everything..
8. Did installing r3.4 ruin all my other package installations? of course it did.
9. OK, `lodown` seems my first solution was the way to go after all, but no worries, a learned a tonne and it was fun.
10. Now I also have to not forget about the makefile… I need to 'touch' all the rds files though, don't want to do them again on mobile data. So I did: `system("touch data/interim/*.rds")`
11. And figured out how to change the direction of the plot see here for more: there's a `rankdir` variable produced by the python code, set to BT, and i can use `sed` to change it inside the makefile.
12. Now makefile seems in order, let's touch what needs to be touched, so I can get on with this. Hm, this touching is not working. I've got a `.touching` file so i can invoce these commands later, but they are all now properly touched in the right order, and `make` is still running..
13. apparently there is a `--just-print` option for make ;). Very useful. SO first of all 02-clean is called for all multiple .rds targets, each time, that's not cool.
14. OK, so that was AN HOUR lost because i had a space at the end of a folder variable declaration in the makefile. (NB: there looks like a good make book is available free here: [http://www.oreilly.com/openbook/make3/book/ch02.pdf])

15. there was more to it... the **pattern rule** again, which I'm not sure I understand anymore now.. I ended up using % inside (what I thought were) variable names... But it worked, and stopped running the clean file 145 times. Commit working skeleton. (the Rstudio build tab is throwing an error saying the build directory does not contain a makefile when it does?).
16. But it didn't really, shit.
17. I'll figure it out later, let's try and figure out why the processed files didn't have labels. Rerunning `02-clean.R`
18. OK, plotting. SO `eulerr` is beautiful. But if you only have two circles, the fitting is presumably easy, but it produces plots where the centres are at a different angle each time. Which makes it impossible to compare visually.
19. but the `fit` gives me the cetroids and the radius, which is all I need, right? SO if I get the distance from the centroids from there, and put it at an angle, the same angle every time, I can plot this in `base r`?
20. Man, `eulerr` ones are beautiful, got them working with region, but need to place them side by side at same orientation.
21. This will be great. OK, got the base version working, but `plotrix::draw.circle` doesnt do transparencies, so I'll have to rewrite that function I guess. Density lines just aren't as good.
22. or use `rgb()` which has alpha ;). but check [https://github.com/jolars/qualpalr] for pretty colors

Friday 22.12.2017

1. OK, let's chech out the `qualpalr` colors. They're colour blind friendly, look pretty, but I also need to find a good combo that will produce a good overlay distinction.
2. Clean up the `FunBasePlot` code. The funciton should be self-sufficient - as much as possible i.e. I've added the package:: calls to the functions instead of loading the whole package when the funciton is called. That way I can call the funciton from another script later and not worry about where the funcitons are from.
3. Something weird with `rgb()`, won't lett me pass a vector, only individual r, g and b values... works if you pass them as a matrix, but then alpha doesn't work. BUG?
4. ARGH, the euler plots are pretty, but perfeclty useless.. I'll have to switch back to barplots...

Wednesday 10.1.2018

1. OK, i'll start thinking about the layout
2. Email Keith of DHS to see wtf is the Senegal deal anyway.
3. it seems some surveys don't have both urban and rural (second 2007 dominican one for example.) Check which ones and remove? Yeah, that's the only one. It's sugar cane plantation workers, I'll take it out.
4. Now I need to add the regions to the catalog, would make it easier to do the layout. Found a UNcodes file from two factsheets back, i'll use that. But have to manually change some names...

Sunday 14.1.2018

1. Get regions sorted out, this should happen at the end of `02-clean.R`. So we're changing the catalog names, not the UN ones.
2. 4 regions: Africa 84, Americas 24, Asia 34 and Europe 3. So let's maybe look at subregions instead.
3. Trying to get more resolution, sub.region name doens't help much, because all of ssa are togheter, but tehre is an intermediate region name for them..

Monday 15.1.2018

1. Hm, it migth be easier if I throw out some repetition e.g. Peru has 7, 2000 - 2012, but they hardly budge. Maybe max them out at 4? Lets have a look at the ones that would have more.
2. Actually, maybe use the phase of the survey to align them somehow horizontally?
3. OK, this looks good, but need to clear out some stuff:

- Egypt has 2 phase 5
- Senegal has 3 phase 7 and 2 phase 6 (continuous DHS?)
- Dominical Republic has 2 phase 5 ones
- Peru has 3 phase 6 and 3 pahse 5 (continuous DHS)

Should I just merge them together? OK, there will have to be a bespoke script in the plotting function to merge the ones here before calculating the proportions… And a note on the factsheet.

4. OK, used excel to manually plan out the layout of the whole thing. in `layout.summary.xlsx`. But now I'll remove the helpers for that, it's not important for reproducibility anyway.
5. OK, now need to add code to merge together the four country/survey combos above..
6. ALso clean up the plot function, throw out the euler ones i'm not using any more..
7. SPlit the plotting funciton and the table prep function, keeping it clean.

Tuesday 16.1.2018

1. OK, so we need to make sure all 5 possible options are always listed, and that the order is 1,3,2, other.. So the `FunTablePrep` function has to make that clear. OK, looks like i can reorder them and ignore the fact that there are sometimes 4 or 5 and sometimes only 3 rows - the NA rownames just get ignored in the plot, so it's fine.
2. Ah, I also want the N for each survey. And the year. And to get the N it's cleaner to remove the NAs from each TablePrep, OK, now it has both.
3. Now make sure `FunBarPlot` uses all three bits of info
4. But wait, no, the NAs from 1 don't always work fine… e.g. no 20. OK, `na.omit`.
5. And actually, all the text rendering will happen in latex anyway, so i don't need to worry about the table prep, instead I need to update the catalog, and use a subset of the catalog later, in latex. OK, so table prep is now really OK.
6. `FunBarplot` should also save each file.
7. Odd things:

- Peru has a wave marked `PE6` in row 106, which is phase 5 according to the catalog. This is 5I. This must be an error, but not important for me atm.
- Egypt also has sth weird, two waves that are both in the same phase, so they got merged here. So I need to fix this and stop it from merging..Actually no, I won't. Keep it simple, one chart per phase.

8. OK, so now I've got all the charts and the psftag table, to help me automate the naming in Latex.
9. But what on earth have I done with the makefile.. I need to make sure this works at some point, but right now probably good idea to move on to LaTeX.
10. OK, last R thing for now, how do i make the psfrag thousands look nice, with commas?
11. Need to get to grips with `multips` as well, this layout stuff manually is shit. e.g. herehttps://tex.stackexchange.com/questio
pstricks-i-need-to-created-the-following-grid-sheet-which-fits-well-with

Wednesday 17.1.18

1. OK, starting laying out in Latex, but it;s all very cramped. No room for N, also tricky doing both ruban/rural, as well as year of survey.. Let's try more white space. And also change the ratio, theye need to be wider so I can get all of them on the plot..
2. Also in pstricks need to use variables, definitions of values to stop copy pasting stuff around. Blog idea, this would be neat to sum up, for my own reference. seems `\def` causes problems if you do e.g. `\def\indent`, and presumably that already exists, so it fucks up and doesn't even give me an error (flashes it too quick for reading.
3. Odd things:

- Ghana 2014 says phase 7, but wave is Gh6 (I put it down as 7)
- Senegal overwrote SN6… need to go back and fix.
- Because same as Ghana, the last three are phase y, but the wave is recorded as SN6.. This I will have to fix manually.
- Ethiopia as well, says phase 5 but wave 4 for the 2005 one.
- And Rwanda, two actually there.

4. Also the axes are booring, I can do gridlines instead.
5. Half way there..

BUT - need woder and lower plots to fit it all in. Probably best thing to do would be to move year to the left side, rotated.

Thursday 18.1.18

1. Rwnanda all four have psfrag split codes, need to maunally do them.
2. Add white background for years.
3. Move latin america right.
4. What;s the deal with the dominican republic? There seems to be an error on the DHS side, why are there two phase 5 files, different id, but same number of cases, 5A and 52? Right, 5A is a special one, BATEYES, sugar plantation workers. Wait, but why are they the same size then? Oh, did my code merge them? Hm, looks like i didn't remove it when i said i did. that's because i'm beign so shit witht he make file, the makefile would have totally prevented any such shit.. OK, fixed that by reruning 02 and 03.

5. OK, now Colombia waves are off as well, one was def overwritten..
6. Wave 4 in Armenia is also wrong. And Bangladesh.
7. Looks like one of the Indias is only Kerala.. hmm. I should make a note of that
8. Also Cambodia has a wrong wave.
9. OK, they are all now in place, now polishing
10. TeX math: the calculation stuff works for ps coordinates, but for other stuff, e.g. the width of a parbox, you need to use the `calc` package. And there you don't define, you use `\setlenght`... Anyway, this is for a how to blog post some day.
11. A bit of rejigging, making sure everything fits tight and has the correct ratios.

Friday 19.1.18

1. Create legend.
2. add gridlines leading in everywhere?
3. Fix text.
4. complete draft.

Monday 22.1.18

OK, so poster is complete, what needs to be done now.

1. I need to make sure the makefile is clean and actually runs.
2. I need to write up a methodology report, short version for teh README, and long version for report.
3. I also need to make the tabled data clean and available, and reflect that in the README.

OK, got the data out. Need to think about the weights. What is the unit of analysis? Women?

Tuesday 23.1.18

1. README: needs structure and pointers to where the important things are. OK
2. Phase VII - did they also ask currently married/non-users? the recode hasn't been published yet. Looks like they didn't, that's a different question, so the recode works fine ;) (Make sure you fix the question on the poster)
3. Damn, but the missing values are inconsistent. now i need to find v502 and 312 as well, to get the correct 'universe' of women, and know for sure which ones are missing and which ones not...

Wednesday 24.1.18

OK, so update 1: remove all the 'missing answer' ones. That is the quick, first option. Although it would be more informationful to know the missing answers... Should I go for that option instead?

1. OK, first fix FunDataExtractor to add the `v312` and `v502` variables.

2. then rerun `02-clean.R`.
3. OK, this seems to have worked, and hardly changed anything except for Kenya and Turkey, which now have around half missing!? Hmm, weird.
4. move the final.csv to human.readable.resutls
5. OK, this is now publishable, only the makefile/reproducibility issue needs to be fixed.

ToDO

- graphviz prettify
- test minimal makefile example before making this whole thing work again