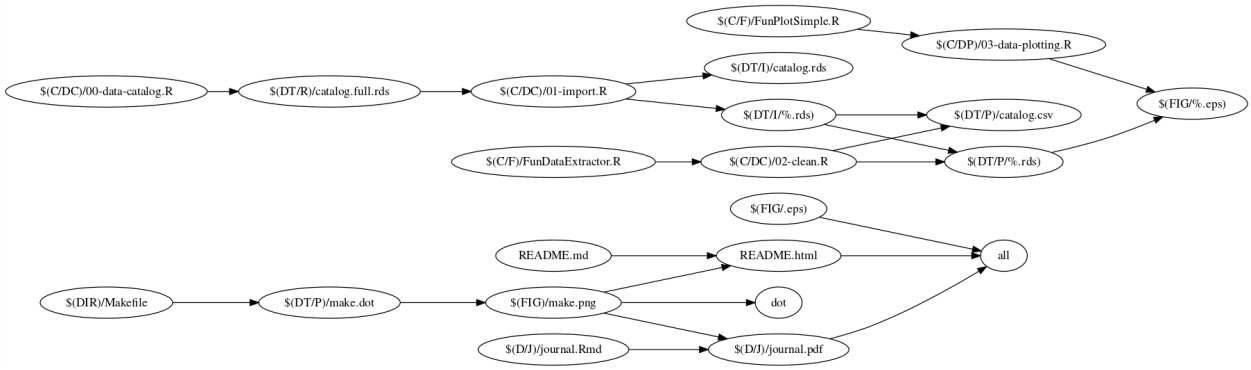
Journal

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?)

ToDO

- graphviz prettify