Monday 13.8.2018

1. OK, basic outline copied from original factsheet, now needs cleaning up.
2. Renamed the data folders and couldn't get `make dot` to work, turns out that's because the `dot2png.R` file had hardcoded paths. This needs to be changed:

- The makefile must pass the paths to the file. How do I do that again?
- Actually this question is about how to pass arguments from the command line. Let's try this source.
- OK, so using `commandAgrs` in R works, it picks up anything after the `Rscripts` command, and makes it avaylable in R.
- Now the only problem I have is that the path I am trying to pass has a `./` at its start in the makefile, but not in R. Used string substitution: `$(string:pattern=replacement)` so `$(DT03:./%=%)` to remove the `./`

3. Now streamline the makefile to get rid of all the factsheet stuff. I'll remove the journal as well.
4. Also don't forget to include the fact that you know now that multiple targets do not lead to multiple executions of the rules, only multiple printouts using the `-n` option.
5. 01-import OK, run to interim, so that downloads and saved data to interim are done. This takes a while, since I't downloading all 250MB of data.
6. In the meantime looking at 02_transform. I am not clear why I use the `FunSpline` function when the spline function seems to do what I want?

Tuesday 14.8.18

Notes on publication:

*Research Data Journal for the Humanities and Social Sciences might be intersting, although it seems pretty new and not really doing such niche things.

*Scientific Data with Nature looks great, but you need to pay, plus it is aimed at bioinformatics and the like. But good template "Data Descriptor" - see also in `\literature` folder for template.

*Open Health Data might actually be best bet. With an excellent example of some guy who cleaned up the WHO mortality database (a bit) and made it available here

*Data is also a data journal, but not social sciences based, although there was an interesting call for papers on grid population datasets, so keep that in mind!

OK, so where was I?

1. `.gitignore`
2. What was the difference between FunSPline and spline? OK, so it seems that `spline()` cannot take the "mono" method. And I want to use that one, since it means the function is monotonically increasing, and also since the data fit the IIASA factsheet. That is for calculating the threshold.
3. So I will best use the "mono" function for the rest of the interpolation as well! OK
4. See if I can't get 1950-2100 data instead of 1953-2098. OK
5. SO having added the extra years at the start and end, the mono funciton makes a big difference there especially. So yeah, FunSPline all the way.
6. Next step: getting the populations sizes at the threshold age: I only did that for both genders together, but it would make sense to do it for them separately as well?
7. First I'll just go back and clear up the `pop` dataset - it has too many variables i don't need.

8. Now split the population size exptrapolation into three threads, one for each group. (Could probably be done in a more modular way, but this is too non-standard evaluation teritory for my lady brain).
9. OK, popoulations at threshold age are now caclulated, now just find the ones over or under.
10. Double check with the MENA data for both genders together - looks good (at least in Algeria).
11. Careful, the demo data files might need following up since i've changed the variable names.
12. In joining three tables the use of suffix is not great - it only works if the variable names are the same, but if you've already joined two tables, then they aren't any more, and it doesn't force the suffix.
13. LEQ 65 or < 65?!?!?!

Wednesday 15.8.2018

1. OK, so clearly we are intersted in people aged 65 and over, so my previous versions were wrong, since the under were LEQ 65 instead. OK, and this has now been double checked with World Bank data which is based on the 2017 WPP, so the calculations done now are correct.
2. So how does the makefile look now?? the `interim` was just temporary, wasn't it?
3. OK, fix up `methods.Rmd` in line with the cahnegs to 02_transofrm from yesterday, esp. spline function and variable names.
4. Oh man, cleaning makefiles - it's really tricky since there are no errors if e.g. a filename i named doesn't actually exist.. So i renamed a few files and didn't update the makefile, and that takes ages to figure out..
5. OK, methods are OK.
6. Now this is the repo for the prospective age data, so no need to save the population data. But because I need those int eh factsheet I need to move that part of the code there.
7. And in the factsheet have to now import the `prospective_age` data instead of transforming it there. Also, there won't be any methods file there any more!! Or rather it is only compiled here, in the data repo, but I could link to it there as well.
8. OK, tried to clear out the data transform code in the spreadsheets and turns out i got the zigzaggy proportions over the threshold again… Apparently that's because of the <= threshold! Fixed.
9. Rerun whole factsheet, fingers crossed. All good. Commit.
10. Still has a manual copy of the data, I'll download it from figshare as soon as I deposit this there for the first time.
11. What next. I need a codebook?
12. But first, set up the github/figshare setup! So here are some things:

- i can add a github repo as an "item" on figshare and set it to automatically update at each new release.
- aha, so on github I can have releases!
- but then it's the whole repo that has a DOI
- what if i want a fileset with the data, the codebook and the methods instead.

13. OK, actually, I really hate the use of capitalisation in the variable names, so that needs to be cleaned up! Done that, but that will have reprecussions in the factsheet code as well.. I think that's all sorted now. Factsheet works great. Commit now.
14. OK, so codebook.
15. ToDO: region identifiers

16. ToDo: other measures?

Thursday 16.8.18

1. POlish up the methods a bit, make it presentable.
2. Tidy up dataset! Yeah, i think it will be nicer to have location/time/gender rows instead of the wide format i have now. It will require a spread in the factsheet analysis. OK, both sorted.
3. Now make sure the dataset is somewhere online and download it directly in the factsheet. I want to be able to cite the dataset with metadata on the factsheet. Hmm, maybe `rfigshare` is what i need now. Trying this tutorial here.
4. Tutorial is pretty deprecated… I get it to work, but authentication happens via a browser, so I can't automate it - at least not obviously how.
5. Man, the `rfigshare` path was not simple at all. But I got it in the end. Could totally do with a tutorial though!
6. So how might someone else point to the most recent version of the datafile? Programatically? Because the file id changes, but the name of the file stays the same.
7. Sorted

Friday 17.8.2018

So plotting the male and female proportions along with the normal proportions over 65 and over the threhsolds revealed some issues, first of all with Tunisia.

In trying to sort it out, I figured out sth else: My proportion over the threshold was calculated using the total population cumsum and total population old-age threshold. But men and women have separate thresholds, so one could argue that they should be calculated separately and then added together to get the proportion over the(ir) old age threshold. Still haven't solved the second part though.

Monday 20.8.2018

1. But also actually this is not necessary that the total is a sum of individual props over threshold, it could just be the sum of people over the total trheshold, i just need to clearly state that.
2. And now back to fiugring out what the problem is with Tunisia..
3. I think I figured it out. The problem is when the threshold is just under 65, then the threshold and value gets picked as the last value under 65 instead of 64! SO Instead I'll just switch to <= 64 instead of <65.
4. UGH, but no, it means <= 65 was correct to begin with, because these are cumulative sums of populaiton, so <=65 in this table is the cumulative sume up until 65. So I had it right before. But that means the world bank data is wrong?!
5. Let's switch that back now. And then when i'm online check if I can't find any UN data to compare to, and double check the oer 65 proportions I have are right.

Tuesday 21.8.18

1. OK, draw it, step by step. CumSum at line age == 64 is the total population of people aged 64 up until almost 65. So <= 64 is correct to get the proportion over and under 65.
2. The problem is in the interpolated CumSum at the threshold. Because that's a point estimate, not a agegroup (albeit single year) estimate.
3. So AgeGrp was the orignal variable name, renamed to `age` here, but it's really age group start, which means `agegroupend` is age+1, and I need the end of the age group for the cumulative proportions to be the point estimate. So i'll add `age_group_end` to be nice and clear here.
4. And just a note for fun, half way through this project i started converting to underscore variable naming instead of periods, ha. So everything is naturally a bit inconsistent, soz.

Thursday 23.8.2018

1. OK, finish rewriting the splines for cumulative populaiton correctly. I think that's all good now.
2. Now let's see how the plotting funcitons work with these new datasets.
3. But first actually have to run the makefile here! OK; all almost worked up until publish, since i didnt have rfigshare installed here.
4. Excellent, this works! Back to the factsheet now.

Tuesday 28.8.2018

OK, so the splines are now fixed and presumably correct. I need to double check the numbers where i can and update the codebook and the methodology paper.

1. Comparing my results with the Wittgenstein centre Data Explorer is not a good idea, since that seems to be based on the 2014.
2. Compare with world bank data directly. All correct, see `04_double-check.R` for code.
3. Clean up codebook to account for all the checks and how correct everything is.
4. methods charts are missing axis labels
5. Looks like i overwrote a new version of the data transform when i got back from France.. Because the data is wide. So used dropbox to get the old version back, hopefully it's correct! Hmm, when did i go wide? The codebook was wide first and then rewritten to long. All the factsheet stuff is clearly written for wide. So I could just fix the codebook back.

Monday 12.11.2018

Started preparing the prospective ageing talk and noticed the figsgare most recent version - 19 - doesn't have the data, just the manual and codebook!? what gives!

2. OK, so my code is not safe, not sure how i lost the file, but the update looks up the filenumbers of the newest version, and so with one missing, that meant it wouldn't pick it up next time either. .