Onboarding to work with register data
DARTER Kickstarter

Register Space Meeting, September 26 2024

Agenda

- Challenge: How to work efficiently on DST
- Potential solutions
- DARTER Kickstarter
- Current and future challenges
- Experiences: Ida & Vithiya
  - Others?

Challenge

- Closed environment
- No introduction provided by DST
- Workflow is **not** self-explanatory
- Messy folder/file structure and naming
  - R.I.P. non-native Danish speakers

Solution

- Don’t reinvent the wheel
  - Share & reuse knowledge/code
- Shared challenge & workspace:
  - Potential for shared guides, utilities & workflows

DARTER Kickstarter

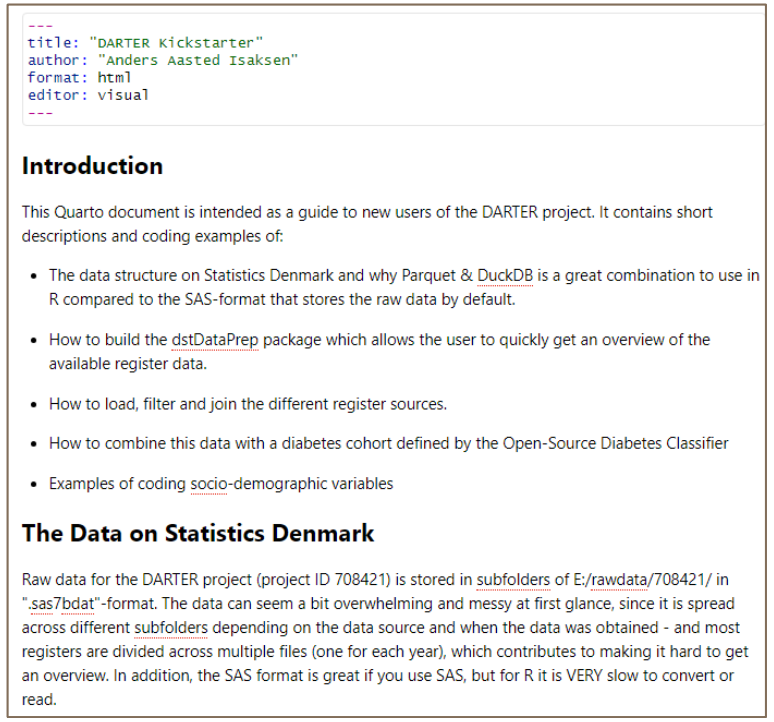
- Ad-hoc onboarding resource
- Most new users will be using R
- Luke's R package: dstDataPrep
  - Handles "folder hell" in raw data
  - Processes data extremely fast: *Parquet/DuckDB*
  - Standard *R/dplyr* syntax
    - A few twists (database connection vs. in-memory)

Current format

Quarto document

Text and executable code

E:/workdata/708421/
workspaces/_onboarding/
darter_kickstarter.qmd



```
---
title: "DARTER Kickstarter"
author: "Anders Aasted Isaksen"
format: html
editor: visual
---
```

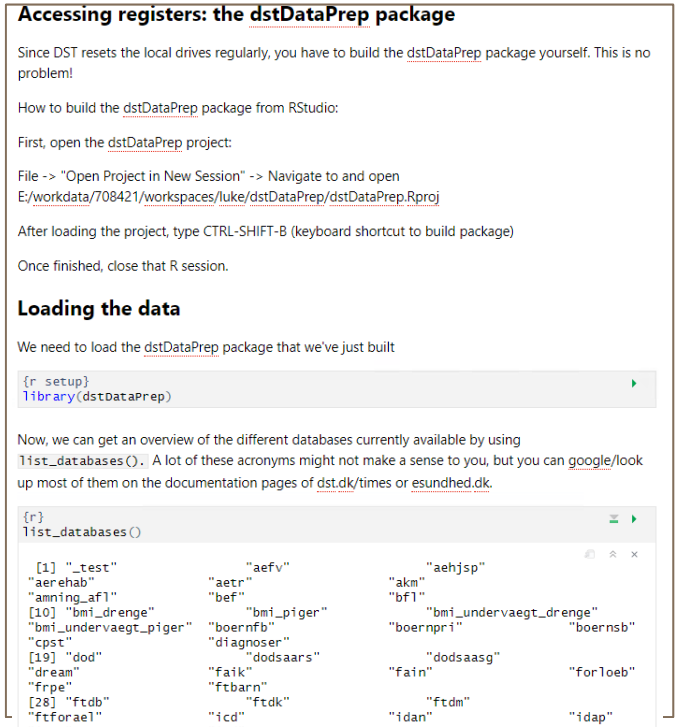
Introduction

This Quarto document is intended as a guide to new users of the DARTER project. It contains short descriptions and coding examples of:

- The data structure on Statistics Denmark and why Parquet & DuckDB is a great combination to use in R compared to the SAS-format that stores the raw data by default.
- How to build the dstDataPrep package which allows the user to quickly get an overview of the available register data.
- How to load, filter and join the different register sources.
- How to combine this data with a diabetes cohort defined by the Open-Source Diabetes Classifier
- Examples of coding socio-demographic variables

The Data on Statistics Denmark

Raw data for the DARTER project (project ID 708421) is stored in subfolders of E:/rawdata/708421/ in ".sas7bdat"-format. The data can seem a bit overwhelming and messy at first glance, since it is spread across different subfolders depending on the data source and when the data was obtained - and most registers are divided across multiple files (one for each year), which contributes to making it hard to get an overview. In addition, the SAS format is great if you use SAS, but for R it is VERY slow to convert or read.



Accessing registers: the dstDataPrep package

Since DST resets the local drives regularly, you have to build the dstDataPrep package yourself. This is no problem!

How to build the dstDataPrep package from RStudio:

First, open the dstDataPrep project:

File -> "Open Project in New Session" -> Navigate to and open E:/workdata/708421/workspaces/luke/dstDataPrep/dstDataPrep.Rproj

After loading the project, type CTRL-SHIFT-B (keyboard shortcut to build package)

Once finished, close that R session.

Loading the data

We need to load the dstDataPrep package that we've just built

```
{r setup}
library(dstDataPrep)
```

Now, we can get an overview of the different databases currently available by using `list_databases()`. A lot of these acronyms might not make a sense to you, but you can google/look up most of them on the documentation pages of dst.dk/times or esundhed.dk.

```
{r}
list_databases()
```

```
 [1] "_test"                  "aefv"                   "aehjsp"
"aerehab"                    "aetr"                   "akm"
"amning_afl"                 "bef"                    "bfl"
[10] "bmi_drenge"                 "bmi_piger"              "bmi_undervaegt_drenge"
"bmi_undervaegt_piger"       "boernfb"                "boernpri"               "boernsb"
"cpst"                       "diagnoser"
[19] "dod"                        "dodsaars"               "dodsaasg"
"dream"                      "faik"                   "fain"                   "forloeb"
"frpe"                       "ftbarn"
[28] "ftdb"                       "ftdk"                   "ftdm"
"ftforael"                   "icd"                    "idan"                   "idap"
"iepe"                       "ind"
[37] "koder"                      "kontakter"              "kotre"
"kronikere"                  "lab_dm_forsker"         "lab_forsker"
"lab_labidcodes"             "lab_optaelling"         "lmdb"
[46] "lpr_adm"                    "lpr_bes"                "lpr_diag"
"lpr_sksopr"                 "lpr_sksube"             "lprmfrdf"               "lprmfrlf"
"mfr"                        "maalinger_afl"
[55] "nyfoedte"                   "ophold_plejehjem"       "organisationer"
"pop_case"                   "preschooldata_mf_dst"   "procedurer_andre"
"procedurer_kirurgi"         "psyk_takst"             "resultater"
[64] "rygning_afl"                "shss"                   "sogn"
"soma_l"                     "soma_t"                 "soma_takst"             "ssnv"
"sssy"                       "sysi"
[73] "t_grund"                    "t_ivf"                  "t_lfoed"
"t_psyk_adm"                 "t_psyk_diag"            "t_tumor"                "udda"
"udfk"                       "udg"
```

Current contents

- Working with dstDataPrep:
  - Listing and loading register data
  - Selecting variables, filtering rows, joining tables
- Using the OSDC cohort:
  - Loading, joining to background population, filtering to prevalent cases

WIP contents

- Adding sociodemographic data
  - Personal income/household income
  - Formatting sociodemographic data
    - what in the world is
    - ”n_audd_pria_l1lx_k.sas7bdat”?
    - ”n_socio_13_kt.sas7bdat”?

Possible future content

- Common functions for common use-cases
  - source(helper_functions.R)
  - add_variables(

    population = my_pop,

    index_date = ”01-01-2018”,

    variables = c(”3y_income”, ”education”, ”employment”, ”diabetes”, ”prevalent_cvd”, ”incident_cvd”)

)
  - Project setup, Git use?
- Needs? Suggestions? (discussion for later)

Challenges

- Current:
  - Individual:
    - Learn programming/statistical concepts
    - Learn R syntax/workflows & DST/*dstDataPrep* niches
- Future:
  - Organization:
    - Package & document maintenance/updates
    - **Project database: Workspace fragmentation**

Experiences

- Ida & Vithiya
  - Intended users:
    - Junior researchers, non-technical backgrounds
    - Non-technical supervisor group
    - No previous experience working on DST

- Others' experiences?
- Needs? Suggestions? Non-R tools?