R on DST

Tips for efficient use of R on Statistics Denmark

Christian Torp-Pedersen Kathrine Kold Sørensen Amalie Lykkemark Møller

06 January 2022

Preamble

This document is made with the purpose of aiding use of R on servers in Statistics Denmark. We will not exclude that it can be helpful elsewhere. The particular characteristic of this environment is the very large size of datasets. For this reason data management is in general suggested to use functions from the package "data.table" which is specifically designed to handle very large datasets. Most introductions to R refer to "data.frame" and "data.table" is a data.frame with some add-ons. There are plenty of pdf and introductions to data.table on the web. This document does not replace such reading. One good place to get an overview of R is http:/r.sund.ku.dk. This introduction is focused on data.frame for management.

The structure of this document attempts to follow the typical path of a project: Include data, manage data and finally perform calculations and produce graphics.

Selected chores make use of functions in the package "heaven" which is made specifically for use in Statistics Denmark. This package is available on github and can be installed with

```
devtools::install_github("tagteam/heaven").
```

Statistics Denmark is a closed environment and all packages available are pre-installed, therefore only library statements are necessary. All packages on CRAN installed on the servers used by us.

In general there are multiple ways of reaching the target of a project. The suggestions in this document are not rules to be followed, but reflect experience after many years of working in the particular environment. Suggestions for improvement are always welcome.

Most R programs start with a list of **library()** statements to provide connection with functions in selected packages. Another possibility is to provide both the package name and the function name for a call. In this document we will try consitently to use the latter convention to show which packages needs to be included in a program. For standard use library statements should be the rule. The function 'thisFunction()' in 'thisPackage' is in this document shown as **thisPackage::thisFunction()**. The exception to that rule is the package **data.table** where many functions cannot be shown indirectly.

Housekeeping rules

In general there are multiple ways of reaching the target of a project. The suggestions in this document are not rules to be followed, but reflect experience after many years of working in the particular environment. Suggestions for improvement are always welcome.

Most R programs start with a list of **library()** statements to provide connection with functions in selected packages. Another possibility is to provide both the package name and the function name for a call. In this document we will try consitently to use the latter convention to show which packages needs to be included in a program. For standard use library statements should be the rule. The function thisFunction() in thisPackage is in this document shown as **thisPackage::thisFunction()**. The exception to that rule is the package

data.table where It is important to structure the data of a project in such a way that a logic is maintained - and also a logic which can be identified by others. In the common circumstance that help is required by others this logic is tested. It should also be noted that a scientific project can later be criticized and require a need to document how a result was obtained. In this situation a logical structure of the project is necessary when the programs need to be rerun years later.

For a beginner it is tempting to produce multiple programs as a project is developed mixing data management with analysis which can easily result in a situation where the project contains a large block of a mixture of outdated and current programs - where the content of each program needs to be rerun in a specific order to obtain the results of the project. Therefore, users are strongly encouraged to follow certain rules.

- Create a directory for each logical part of a project - in general a directory for each paper. A directory is essentially the folder you are working in and will be the folder you are saving figures, datasets, table etc. in. We recommend that this directory is created using **heaven::createProject("path to project")**
- Separate data management from analysis. Often it is useful to have two files for a project, one for data management and the other containing the analyses and statements that produce the tables and graphics for a paper.

- Keep the number of programs very low and format them to follow the same logic as the paper you are writing. This should not prevent you from having multiple versions that are numbered such that the final version is easily identified. An alternative is to use git within Statistics Denmark to keep track of versions.
- Comment VERY generously. You will be surprised how short memory is regarding variable names etc.
- During development of a program many little tests are performed. Remove these statements when testing has been done. If you want to keep testing versions of a program use a separate directory such as the "sandbox".
- When you make major changes, then keep the old version either numbered or in a selected directory. If your results suddenly change this can be the only way of determining whether the changes reflect correcting a problem or creating a problem.
- Keep only moderate sized analysis datasets as files. Huge datasets can be created and clog the resources of our environment. Old datasets can be recreated by rerunning programs.
- Keep the final analysis dataset until the paper is published. Data on Statistics Denmark are periodically updated which can result in changes.
- All datasets should be deleted when the paper has been published, but you need to make sure that the programs can be found even years later. In line with this requirement is is unwise to have programming which reads derived data from other users. If you need such data, then copy the programming that generated the data to your folder to make sure that your results can be derived with your programs directly from the raw data provided by Statistics Denmark.
- It is wise to request export of programs so you can make sure these are not changed or deleted by mistake. To enable this without risk it is critical that programs never include statements that involve "microdata". An example of such a problem can be a logical statement that involves a person identifier variable.
- We encourage that the total output to produce a paper is programmed. Most will benefit from "R markdown" and geeks may choose EMACS-ORG. The advantage comes as a project develops. It is very common that the tables and figures for a project needs to be generated many times pending on various decisions regarding data. Having a program that generates everything is in such cases a great advantage. These output programs should not include very time consuming calculations. Therefore it is often advantageous to have one program generate the data as R objects and then have another R markdown program produce the output.
- Keep programming compact! It is common in programs that similar chores are repeated for slightly changing conditions. It is tempting for the beginner to copy-paste programming sections several times and make slight changes in each block. This makes programs difficult to read and difficult to maintain. Whenever such chores are needed the repeated blocks should be replaced by functions and/or loops.
- Make programs where not everything has to be rerun after changes. For many projects a single final dataset is used for many calculations. This can be generated by consecutive addition of more and more

variables from other datasets. In this case if just one of the steps require a change everything has to be rerun. It is wiser to create the components independently and then have a final large merge step in the end. In this case only the component needing change and the final merge needs to be rerun.

Memory use

R keeps all data in RAM (random access memory) and the program is prone to "memory leakage" which implies that large chunks of memory are blocked from all users, but do not contain usable data. It is common to include very large datasets during data management and once these are not needed anymore they should be removed with **rm(file1,file2,file3)** (names of data separated with comma).

It is important to realize that this command does not free the RAM you have used, so when chunks of data have been "removed" you can clear the memory use with the garbage collector: **gc()**

The Rstudio program can under the pane "Environment" show the use of RAM memory. The garbage collector also reports on memory use and the operating system can provide lists of users along with their memory use. This is provided with the **Task Manager**. This feature is the one to use if you cannot undertand why there is not enough room for you or the server appears slow.

Hard drive memory should also be used sparingly, although less so than RAM. What should not be done is to keep long lists of files that represent the history of you data development. Hard drive memory is interrogated with the windows explorer.

Starting a project

If you have used the **create_project()** function to define the project there are suggestions for R scripts that encourage a starting comment defining the author and the purpose of the program.

At start you should use the function **setwd("path to data")** to the directory for your data - or to the directory for your project. This makes reference to datasets in your project easier. If you choose the directory of the project you can use relative paths to read and write in subfolders of your project such as **writeRDS(R-object-name,file="/data/filename.rds")** to write in the "data" subfolder.

Read data

Data in Statistics Denmark is generally provided as SAS datasets - and a few data are provided in ASCII format which includes the very useful "csv" format.

If you need to read an entire table from SAS you can use the **haven::read_sas()** function. The disadvantage of this is that it can take a long time to read a large dataset and the use of memory will be extensive. On the other hand this function can handle nearly any complicated SAS dataset which is not always the case with the import_SAS function described below.

To include data from SAS with limited time and resources we have developed the **heaven::import_SAS()** function that can accept a range of SAS commands to read only a part of the data. The saving of time can be dramatic. The help page for this function provides the many available options - a few are described here:

- obs=n - Read only the first n records. This is very useful during start to ensure that reading is correct while not waiting for the entire dataset to be read.
- keep=c("var1","var2",...) - limits the variables that are read
- date.vars=c("var1","var2",...) - converts SAS date vars to proper R dates. Note that in principle a SAS date is an integer representing the number of days since 01-JAN-1960 and in R it is the number of days since 01-JAN-1970.
- character.vars=c(....) - forces variables to be character
- where="SAS logical statement" - Reads only the records corresponding to the logic. Note that the statement needs to be correct SAS.

- filter=object - if object is a data.table with a single column containing person identifier, then only records from thos people are kept. This is useful if you at the start of a project can define the population and then ensure that you only read records for thos persons from other datasets. *set.hook="SAS statements" - this allows inclusion of such SAS statements that can appear in parenthesis after names of datasets in a SAS data statement. Keep statements should not be included here as they result in errors because the function attempts to define formats for all data except those in the dedicated keep statement.

Since the time to read large SAS datasets very often requires patience, it is usually wise to limit the number of times a SAS dataset is read. Often it is wiser to grab all the data you need in one pass and afterwards use R to select further data for various purposes. Typically you may need medications and diagnoses to define very separate structures in the project - but try to collect all in one pass anyhow.

When you need to import ASCII data (text files) the most efficient function is **data.table::fread()**. This function can usually identify the type of data from the extension to the data file. The help page provides a long list of options for reading.

For most of the programming suggested here data need to be **data.table** rather than **data.frame**. When a data.table function does not work, it is likely that the format was **data.frame** and the most convenient way to convert is the function **data.table::setDT(object_name)**. This function can also be used just to make sure an object is **data.table**. For the rare situation that you need to convert the other way to data.frame the function **data.table::setDF()** does the job.

If an object has been saved as ".rdata" the way to read it is **load("path to data")**. The name of the object is also read and will appear in the environment.

Many datasets are saved as "rds" files, single R objects. They are read and added to the environment with **MyData <- readRDS("path to file")**.

Save data

The most efficient way to store datasets is as single object using **writeRDS(name_of_object,file="PathToFile.rds")**. If your working directory is the project directory and you want the object saved in the subdirectory "data" then use **writeRDS(name_of_object,file="/data/PathToFile.rds")**.

If you need to output text files (ASCII) then use **data.table::fwrite(name_of_object,file="pathToFile.csv")**. This is useful if you want to use Excel and similar to produce tables. Most table are generated in programs as data.frame or data.table and after converting to a csv-object they can be read with Excel. Note that the **fwrite** function senses the details of output format from the file name you provide.

View your data

As data is imported to the project it is necessary to check that what you have desired has actually been accomplished. You can click on object in the environment pane to show the datasets as a spread sheat. This is discouraged since large datasets take a long time to load. Very useful functions are:

- names(object) - display the names of a data.table (or data.frame)
- str(object) - provides a list of variables and the first values
- head(object) - provides the first 5 lines of the data
- object - if a data.table then the first 5 and last 5 are shown - and all if there are less han 100
- object[1:30] - shows the first 30 lines
- object[1:10,.(var1,var2,var2)] - the first 15 lines and only the selected variables
- View(object[1:10,.(var1,var2,var2)]) - can be used as a more conscious alternative to clicking on the object

It is generally useful to create small tables to check that your data are what you expect them to be and a very useful procedure is

- **object[,.N,keyby=var1]** which will tabulate the number of records by var1. Tabulation by multiple variables can be done by replacing keyby=var1 with character vector: **keyby=c("var1","var2"...)** or list **keyby=.(var1,var2...)**.

Manipulating data

This is described using data.table. In this description **DT** represents your data.table object. The general format for manipulating is **DT[i,j,by]**. This is most easily learned by remembering that records or rows can be subset using "i", columns can be selected, manipulated, or calculated using "j", and finally grouped using "by". So the general rule is **DT[where,what,grouping]**.

Data.table attemps to limit use of RAM by not making copies of data. If you write the statement **DT2 <- DT** you will not get a copy of the data but DT2 and DT will address the same data.table. If you really need a copy to manipulate the solution is **DT2 <- copy(DT)**. The avoidance of copying also makes it important to realize when you need to use "<-" to change the data. For the examples below to create a new variable no "<-" is present. Only the new vector with an additional variable is created without copying the rest of the data. If the whole dataset is changed by some command the "<-" is absolutely necessary.

Creating new variables

The creation of new variables will be made using 'j': **DT[,j,]**. A simple new variable examplified by age is created with **DT[,age:=(date-birthdate)/365.25]**. This assumes that both date and birthdate are R dates.

A common target is to make a variable that is the largest or smallest non-missing of several other variables. This is accomplished with **DT[,max:=pmax(var1,var2,var3,na.rm=TRUE)]**. Beware of the "p" in "pmax" this p dictates that the comparison is made on a record level.

Several variables can be created in one step: **DT[,':='(var1=var3+var4,var5=var2*5)]**

When You create variables You need to make decisions regarding the variable type. Commonly used variable types include **numeric, integer, factor, and character**.

- The age example above is a "date difference" and if You require it to ba a numeric you need to enclose the right hand side of the equation in the function **as.numeric()**
- Numerics are "real" numbers with high but limited precision which you can realise by showing that **round(0.5)** is zero rather than one. Therefore it can be wise to use **integer** for numbers that are integers.

- Integers are provided by putting "L" after the number, 25L is the integer 25. Similar to numeric, **as.integer()** converts the variable to an integer.
- A logical statement produces a boolean, either TRUE or FALSE. It is a common convention to use the numerics 0 and 1 for FALSE and TRUE which can be achieved either by the function **as.numeric(boolean.var)** or by multiplying the boolean variable with one **DT[,numeric.var:=boolean.var*1]**.
- Variables that represent classes are **factors** in R. If we assume that sex has been coded as 0/1 you can generate a factor with relevant names with **DT[,sex:=factor(levels=c(0,1),labels=c("female","male"))]**. For practical programming it is common to change multiple variables to factors. A shortcut for this is the function **Publish::lazyFactorCoding(name_of_data)**. This function creates programming lines in the console similar to the sex example for all variables with a selected maximum number of levels. This block of lines can then be copied to the program and make it easy to generate multiple factors.
- It is tempting to change all class variables to factors, but be careful. Many regression functions require the outcome variable to be numeric.

Numeric to factor

In the example above with sex the numeric variable already represented levels. When the variable is a continuous numeric, the convinient way of creating a factor is the **cut()** function. A very simplistic example is **DT[,newvar:=cut(oldvar,3)]**. This creates three levels with equal range. A more useful example is creating of a variable which represents quartiles of the old variable: **DT[,quart:=cut(oldvar,breaks=c(quantile(oldvar,probs=seq(0,1,by = 0.25))),labels=c("Q1","Q2","Q3","Q4"))]**

- Importantly, converting a numeric variable containing only 0 and 1 to a factor will result in a factor with levels 1 and 2, where 0 was converted to 1 and 1 converted to 2.

Naming variables

the function **setNames(DT,..)** can be used to name all variables by providing for ".." a character vector with the new names: c("new1","new2"...) - or you can change selected variables by providing two character vectors.

Many imported datasets have a confusing attitude to "case" with mixtures of upper case and lower case names. Changing all variables to lower case can be achieved with **names(DT) <- tolower(names(DT))**

Selecting a subset of variables

The most efficient way to select a subset of variables is **DT2 <- DT[,.(var1,var5,var8)]**. In some circumstances, such as within a function, another version is: **DT2 <- DT[,.SD,.SDcols=c(var1,var5,var8)]**. The variable **.SD** is "subset of data.table" and refers to a group of columns in the dataset.

Subset selection

If a dataset has been sorted by an individual identifier and another variable such as a date, it is common that you want to select just the first occurrence for each individual. This is done with **DT2 <- DT[,.SD[1],by="individual_identifier"]**. If You instead want the last occurrence **[1]** is replace by **[.N]** and if You for some reason want the second occurrence the replacement is **[2]**.

If You want to select a subset based on a logical statement, and efficient way is **DT2 <- DT[age<100]**.

In case you want to search for all records that start with a selected string you can use: **DT2 <- DT[grepl("^A10",var)]** that finds all records where "var" starts with "A10". This is a very simple example of a "regular expression" which is an extremely versatile tool to search for complicated relation. An overview of regular expressions can be found here: https://stringr.tidyverse.org/articles/regular-expressions.html

It is common to require a number of conditions to be defined each by a range of variable content. An example is to search for multiple comorbidites which each are defined by a range of diagnosis codes. A special function **heaven::findCondition** have been developed for this purpose:

- To use this function You first need to create a "named list" of (start of/end of) codes for each condition You wish to find. An example of such a list is **heaven::charlsonCodes**. You can also define another exclusion list. This can be useful if You e.g. want to find "all cancer except non melanoma skin cancer". You can then have "cancer='C' " in the names list of inclusions and in the list of exclusions you have an element "cancer" with the few condition you wish to exclude.
- With the one or two named list generated the **heaven::findCondition** will find all occurrences of you selected condition. The examples on the help page also shows how you can manipulate the result further.

Removing variables

A single variable is removed with **DT[,var1:=NULL]** and a group of variables is removed with **DT[,c("var1","var3"):=NULL]**. Note that no "<-" is necessary, just the vector with the particular

variable is removed.

Sorting data

The functions **setkey(DT,"var")** and **setkeyv(DT,c("var1","var2"))** sorts the data in ascending order and creates a key for fast use of the sorted data.

If you need sorting an variable order the functions to use are **setorder()** and **setorderv()** with this example **setorderv(DT, c("A", "B"), c(1, -1))** sorting ascending by A and descending by B.

Conditional manipulation of variable

The standard format for a conditioned change of a variable is **DT[var==value,var:=new_value]** Another option which occasionally is more convinient is to use **fifelse()** which needs three (or four) parameters: first a logical statment, then the value if TRUE then the value if FALSE and finally the optional value to return for NA: **DT[,newvar:=fifelse(oldvar>x,2,1,NA)]**

Change multiple variables

Many chores need to repeated on many variables, an example being changing a numeric value of e.g. 999 to NA. This can be achieved by first creating a character vector of the variables you want changed and then creating a loop to make the changes:

```
vars <- c("var1","var2"...)
for (x in vars) DT[get(x)==999,(x):=NA]
```

Note two peculiarities here. If we had written x== instead of get(x)== data.table would not know whether x was an object or representing the loop variable. get() solves that problem by enforcing evaluation of x. Similarly, if we had written x:=NA we would repeatedly create a variable named 'x' - the parenthesis forces the evaluation in this case.

Append - rbind

Combining several data.tables to create one long table can be achived with **rbindlist()** or **rbind()**. If the only input is a vector of data.table objects then names and order of columns need to match perfectly. If one table contains extra variables these columns can be set to NA for the other tables with the option **fill=TRUE**. The option **idcol="file"** adds a column with the name of the object that contributed to each record.

Merge

Combining data.tables by columns as apposed by rows can be achived with **cbind()**. This function simply the first row of each data.table with the first row of other data.tables - and requires that they have the same size.

if two data.tables have been provided the same key with **setkey** or **setkeyv** then **DT1[DT2]** will generate an **right merge**, that is a data.table with those records where DT2 is intact and updated with information from DT1: