R: Core Skills for Reproducible Research

Course Manual with Practical Exercises – Oxford IT Services - 22 June 2016

Maja Založnik

Contents

1 Introduction

1.1 Blurb

This short course covers the core skills required for a budding R user to develop a strong foundation for data analysis in the RStudio environment. Within the framework of a reproducible research workflow we will cover importing and cleaning data, efficient coding practices, writing your own functions and using the powerful `dplyr` data manipulation tools.

1.2 Key Topics

- Reproducible Research
- R Studio and project management
- Importing and cleaning data
- Good coding practices in R
- Standard control structures
- Vectorisation and `apply` functions
- Writing your own functions
- Data manipulation with `dplyr`
- Piping/chaining commands

1.3 Course information

Intended audience Anyone interested in quantitative data analysis using open source tools.

Prior knowledge Knowledge of R (as covered in R: An introduction).

Resources Course handbook

Software RStudio & R 3.1.2

Format Presentation with practical exercises

Where next? Data visualisation: Creating interactive visualisations using R and Shiny course

2 Reproducible Research

Reproducible research means making the data and the code of our analysis available in a way that is sufficient and easy for an independent researcher to recreate our findings.

This is the golden standard of scientific inquiry, and is increasingly and rightly becoming a requirement in academic publishing, and by funding bodies.

It is also a way of establishing better working habits, reduce the potential for error, develop a more streamlined research process, and make for easier collaboration.

Reproducible research does take a bit of upfront investment in learning the tools and setting up your workflow. Luckily RStudio has integrated many of the tools required in one platform, making it easier than ever to apply the principles of reproducibility consistently and comprehensively.

This practical course will focus in particular on how to set-up an RStudio project and associate file and folder structure, and get you on the right track towards literate programming. We will then cover a complete workflow structure including downloading and importing data, *tidying* it up, using some basic programming structures to improve your code, and finally the `dplyr` package, which is already revolutionising data manipulation in R by providing a comprehensive set of tools that follow a very intuitive logic.

There is plenty more that cannot be covered in a 3 hour course. In particular we will not discuss version control (e.g. github) and the *knitting* of text and analysis, or their publishing on-line directly from RStudio. You will be able to see the results of these practices in the way the very course materials are prepared, and they can be accessed on-line in a dedicated repository public github repository: https://github.com/majazaloznik/RepResCoreSkillsR or for extra convenience: http://tinyurl.com/RCSRepRes.

For an excellent and in-depth source on all of this and more, see Christopher Gandrud's book on Reproducible Research in R and RStudio, also in a public github repository: https://github.com/christophergandrud/Rep-Res-Book and for your convenience an old compiled copy of his first edition can be also be found in this course's repository in the `literature` folder.

3 Set-up

3.1 RStudio

RStudio has beyond a doubt become the most popular Integrated Development Environment for R. It is open source and cross-platform. In addition to allowing easy project management and integrated version control, it also provides all the tools for dynamically creating *knitted* documents and directly publishing them on-line. The RStudio crew are also active developers of interactive graphical tools and new developments are constantly being added to an already excellent toolbox. If you prefer to use another environment you should however still be able to apply all of the principles of reproducible research covered here and beyond, although it might take a little bit more work.

3.2 Project management

A crucial requirement for conducting reproducible research, and one that has to be carefully considered before you embark on your analysis, is your plan on how the data, code and outputs will be organised. The project management structure proposed here is just a suggestion, and you should adapt it to your specific needs, but it is highly recommended that you stick to one such system consistently, instead of coming up with 'ad hoc' solutions for every new project.

RStudio makes it extremely easy to divide your work into separate projects, allowing you to neatly organize and access your work. This means assigning a single folder for each project, and within that folder organizing your work into sub-folders. Depending on your type of project this may vary, but a good starting point would be something along the lines of:

Project Folder - data - holds all the raw data files - scripts - holds all the R code, preferable split into smaller, more modular code files - figures - to store outputs of your data analysis, or external figures to be used in reports - outputs - presentations, reports etc. that are compiled dynamically within the project

None of this will matter though, if you do not *document everything* you do! This means never running any code from the command prompt, always writing it into a script file and running it from there. This also means not messing with the original data files, even if they are horribly formatted Excel files, but only extracting the data programatically.

3.3 Human readability

One of the often overlooked principles underlying reproducible research is trying to ensure the human readability of as much of your files as possible. This applies to the types of data and other file formats used, as well as your coding style. Making them human readable is one way of trying to future-proof your work.

3.3.1 File Formats

Try to avoid binary file formats. This includes e.g. .xls files, but also R's native workspace file, which is saved with an `.RData` extension. Text, delimited or comma separated values are safe formats that are easily transferable, human readable, and relatively future proof.

3.3.2 Consistent coding style e.g.:

Another key element of human readability is trying to keep to a consistent coding style. This is not always easy, but it pays to get into some good habits while it's still early. Two excellent starting points are

- Google's R Style Guide
- Hadley Wickham's Style Guide

You do not have to follow these to the letter (they do not agree on everything anyway) - but they should give you an idea about what are good rules to determine for yourself and then stick with. At least for the sake of the future you, who will one day have to re-read your messy, uncommented code.

3.3.3 Commenting

The importance of commenting your code can never be understated. Some programmers advocate *self documenting code* – that is code that is self-explanatory and does not require comments – and this may be worthy a goal to aspire to, but in the mean time: comment, comment, and comment some more.

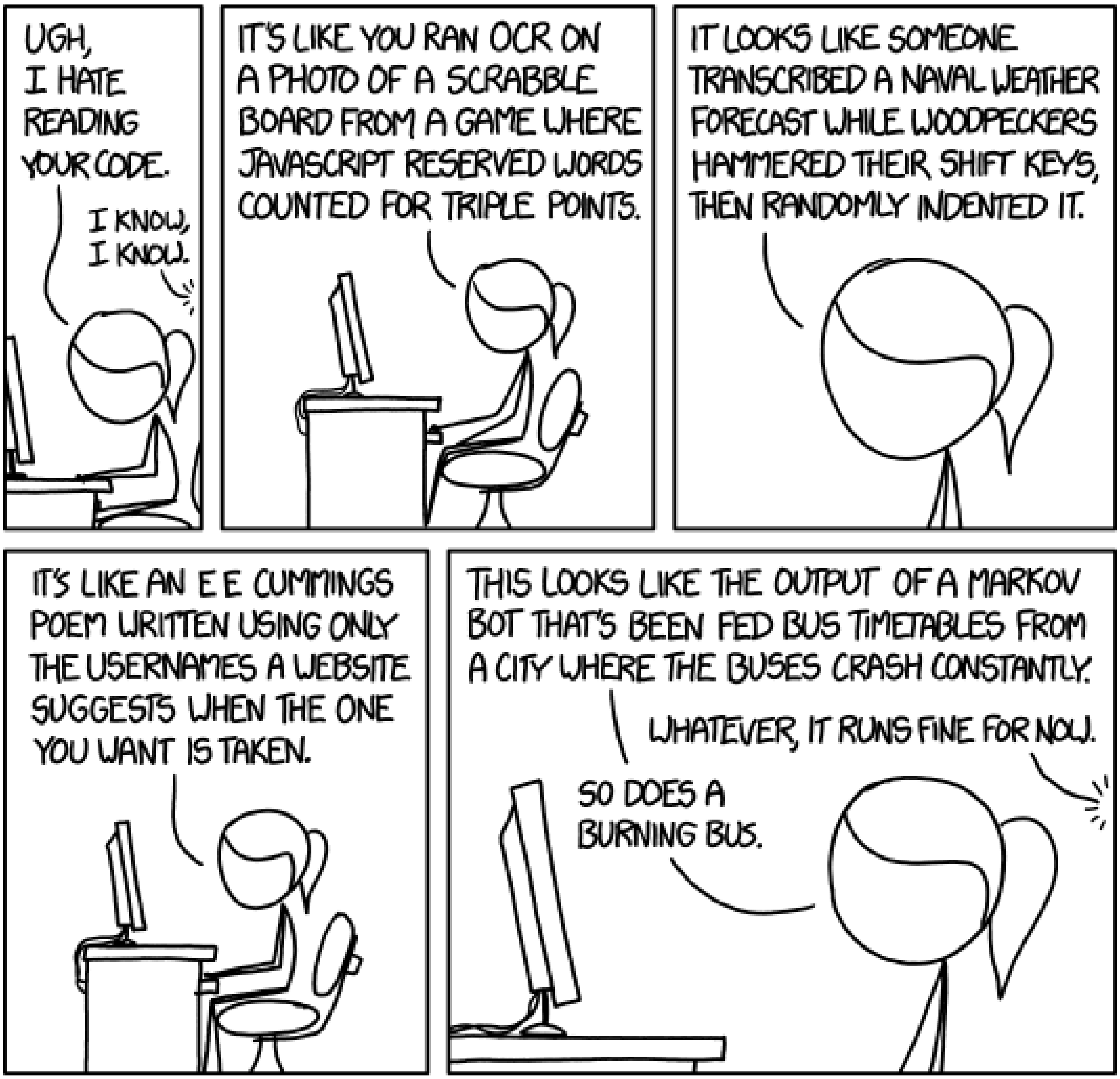


Figure 1: Code Quality part 2. (https://xkcd.com/1695/)

4 Workflow

> *Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.*
>
> source: NY Times

4.1 Importing data

Regardless of whether your data is stored locally or downloaded from the web, you should never manipulate the original data directly. This is crucial for the integrity of your reproducible research process.

R has utilities for importing data from a wide variety of sources including proprietary formats. Ideally you want to be working with .csv files, as they are the cleanest and least problematic to import, but often you have no choice in the matter. In the practical we will import .csv, .xls and .sav files, including downloading and unzipping them. Here is a list of some common formats and the packages used for importing them, refer to the help pages for more details:

- comma separated values – `read.csv()`
- tab-delimited text file – `read.table()`
- other delimited files – `read.delim()`
- Minitab – `read.mtb()` from `library(foreign)`
- SPSS – `read.spss()` from `library(foreign)`
- Stata – `read.dta()` from `library(foreign)`
- Excel – `read.xls()` from `require(gdata)`
- Excel – `loadWorkbook()` from `library(XLConnect)`

The basic import functions of the `read.table()` family all have a `nrows` argument, which is particularly useful if you do not know the structure of the data and are dealing with a large fine. In which case it is recommended you try a test import with e.g. `nrows=10`, and check the result before attempting to import the full file.

For a more comprehensive list of possible input formats see this tutorial: https://www.datacamp.com/community/tutorials/r-data-import-tutorial.

We will store all our data files in the `data` folder of our project, from where they will be imported into R. This means the original files remain *untouched* by the data analysis and should never be overwritten as the result of your analysis.

While you might find it easier to simply download a file into your folder, this poses the problem of loosing track of where the data was sourced from. It is therefore highly recommended you download the data programmatically if possible, and if not, that you use comments within the code to describe the source of the files. For example the `pop2010.csv` file we will download in the first practical should have been downloaded directly from within R, by doing it manually we are reducing the reproducibility of our project. We must therefore make sure we note the origin and date we accessed the data in our code!

4.2 Data tidying

> *Tidy datasets are all alike but every messy dataset is messy in its own way.* – Hadley Wickham

A great deal of data tidying can be done manually with the base R functions. Additionally there are several packages available with more specific functions. In this course we will use the `tidyr` package by Hadley Wickham, which is particularly well integrated with the `dplyr` package we will be using in the second part of this course.

The underlying principle of the `tidyr` package is *tidy data*, which must satisfy the following three principles:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

Source: H. Wickham (2014) *Tidy Data* (available: http://vita.had.co.nz/papers/tidy-data.pdf)

This may seem trivial, but it is in fact common to encounter data that does not conform to these principles. The four workhorse functions of `tidyr` that should solve all your data tidying needs are:

- `spread()`
- `gather()`
- `separate()`
- `unite()`

4.2.1 spread()

Below we can see an example of a messy table, since each observation is in fact represented in two rows. Third column in fact contains variable names (`density` and `population`), while the fourth column contains their values.

```
##     country year        key       value
## 1    Norway 2010 population  4891300.00
## 2    Norway 2010    density       16.07
## 3    Norway 2050 population  6364008.00
## 4    Norway 2050    density       20.91
## 5  Slovenia 2010 population  2003136.00
## 6  Slovenia 2010    density       99.41
## 7  Slovenia 2050 population  1596947.00
## 8  Slovenia 2050    density       79.25
## 9        UK 2010 population 62348447.00
## 10       UK 2010    density      257.71
## 11       UK 2050 population 71153797.00
## 12       UK 2050    density      294.11
```

Using `spread()` we can tidy this layout:

```
tidy.02 <- spread(messy.02, key, value)
tidy.02
```

```
##    country year density population
## 1   Norway 2010   16.07    4891300
## 2   Norway 2050   20.91    6364008
## 3 Slovenia 2010   99.41    2003136
## 4 Slovenia 2050   79.25    1596947
## 5       UK 2010  257.71   62348447
## 6       UK 2050  294.11   71153797
```

The syntax for `spread()` takes the following form:

`spread(data, key, value)`

The *key–value* pair is the underlying logic of the tidy data table. We can decompose the data into a collection of key–value pairs such as this:

Key : Value

```
Country: Norway
Country: Slovenia
Country: UK

Year: 2010
Year: 2050

Population: 4891300
Population: 2003136
...

Density: 16.07489
Density: 20.91484
...
```

In a tidy data table each cell contains a *value* and the *keys* are the column names.

4.2.2 gather()

Here is another messy table:

```
messy.01
```

```
##    country    X2010    X2050
## 1   Norway  4891300  6364008
## 2 Slovenia  2003136  1596947
## 3       UK 62348447 71153797
```

Now we have three variables: the country, which is in the first column, the year, which is across the header row (representing the *keys*), and the population (representing the *values*), which is in the second and third columns. Using `gather()` we can tidy up the table, so that now each of the three variables has its own column, and each row is an observation:

The syntax for `gather()` takes the following form:

`gather(data, key, value, ...)`

where the `...` represents the columns we want to gather, in our case columns 2 and 3. The key and value arguments are the *names* of the two new variables, or columns we are creating: the *key* is currently in the column names of columns two and three - so we want it to become `year`, and the *values* are in the cells of those two columns, so we want it to become `population`.

```
tidy.01 <- gather(messy.01, year, population, 2:3)
tidy.01
```

```
##    country year population
## 1   Norway 2010    4891300
## 2 Slovenia 2010    2003136
## 3       UK 2010   62348447
## 4   Norway 2050    6364008
## 5 Slovenia 2050    1596947
## 6       UK 2050   71153797
```

4.2.3 separate() and unite()

Separate and unite are straightforward helper functions for the reshaping done by gather and spread. The following table for example requires spreading, but the `double.key` variable contains both *values* (years) and *keys* (population and density):

```
##     country      double.key       value
## 1    Norway 2010_population  4891300.00
## 2    Norway    2010_density       16.07
## 3    Norway 2050_population  6364008.00
## 4    Norway    2050_density       20.91
## 5  Slovenia 2010_population  2003136.00
## 6  Slovenia    2010_density       99.41
## 7  Slovenia 2050_population  1596947.00
## 8  Slovenia    2050_density       79.25
## 9        UK 2010_population 62348447.00
## 10       UK    2010_density      257.71
## 11       UK 2050_population 71153797.00
## 12       UK    2050_density      294.11
```

These are separated simply by the following code into `year` and `key`, which can then be used to reshape the table as we did above.

```
tidy.03 <- separate(messy.03, double.key, c("year", "key"))
tidy.03
```

```
##     country year        key       value
## 1    Norway 2010 population  4891300.00
## 2    Norway 2010    density       16.07
## 3    Norway 2050 population  6364008.00
## 4    Norway 2050    density       20.91
## 5  Slovenia 2010 population  2003136.00
## 6  Slovenia 2010    density       99.41
## 7  Slovenia 2050 population  1596947.00
## 8  Slovenia 2050    density       79.25
## 9        UK 2010 population 62348447.00
## 10       UK 2010    density      257.71
## 11       UK 2050 population 71153797.00
## 12       UK 2050    density      294.11
```

The function `unite` is the inverse of `separate`, and merges the values of selected columns into a new single column. In both cases you can change the separator using the `sep=` argument.

```
messy.again <- unite(tidy.03, new.double.key, key, year,  sep = " in the year ")
messy.again
```

```
##     country              new.double.key       value
## 1    Norway population in the year 2010  4891300.00
## 2    Norway    density in the year 2010       16.07
## 3    Norway population in the year 2050  6364008.00
## 4    Norway    density in the year 2050       20.91
## 5  Slovenia population in the year 2010  2003136.00
## 6  Slovenia    density in the year 2010       99.41
## 7  Slovenia population in the year 2050  1596947.00
## 8  Slovenia    density in the year 2050       79.25
## 9        UK population in the year 2010 62348447.00
## 10       UK    density in the year 2010      257.71
```

```
## 11       UK population in the year 2050 71153797.00
## 12       UK    density in the year 2050      294.11
```

For an excellent write-up of the main `tidyr` functions see Garrett Grolemund's post here http://garrettgman.github.io/tidying/.

For a quick tidyr cheat-sheet stick this to your wall: Data Wrangling Cheatsheet, also available in teh `literature` folder of this course's repository.

4.3 PRACTICAL: new R project

The complete documentation for this course is available as an RStudio project in a public github repository: https://github.com/majazaloznik/RepResCoreSkillsR or for extra convenience: http://tinyurl.com/RCSRepRes.

In this first part of the practical you will set up a new RStudio project mirroring the structure of the github repository for this course.

Unless you have brought your own laptop, you will be completing this project on the local drive of the computers here. This unfortunately means you will have to transfer your work via USB, email or other means in order to keep it for your record. The complete materials will however remain available to you on-line at the above addresses.

4.3.1 Create new RStudio project

1. Open RStudio
2. Select the project menu in the top right-hand corner and select `New Project`
3. Select New Directory and choose the name of your project (e.g. RRepResCourse)[1].
4. RStudio has now created a new project file in your folder, which you can see in the Files pane.
5. Run the command `getwd()` in the console. You should note that RStudio has automatically set the working directory in the top level of your project folder.
6. Click on the project menu again and select Project Options. On the options "Restore .RData to the workspace at start-up" and "Save workspace to .RData on exit" select No. In fact, you should set that as a global option (Tools/Global Options) – for truly reproducible research you should never have to load a previous workspace!

4.3.2 Folder and File Structure

1. Using the New Folder button in the Files pane, create three folders called (some equivalent of)

- data
- scripts
- figures
- presentations

2. For the rest of this practical, all you need is to download the file `pop2010.csv` from the github repo into your `data` folder.

- Downloading single files from github can be cumbersome - you need to navigate to the file https://github.com/majazaloznik/RepResCoreSkillsR/blob/master/data/pop2010.csv and right-click on the `Download` button on the top rigth and select `Save link as`.

[1] As a general rule you should avoid spaces in file and folder names, although you will probably be fine if you ignore this advice.

3. You can also download the manual and slides to your `presentation` folder, and the two reference files from the `literature` folder.
4. Create an .RProfile file using the command `file.edit(".Rprofile")`. This will open a new script tab for you to edit. The content of the .RProfile gets automatically run every time you open your project. It is therefore a good idea to e.g. load any packages you will need from your .RProfile, instead of doing it manually every time.

For this practical you will need the following packages (type this into your .RProfile file and save).

```
require(xlsx)
require(foreign)
require(dplyr)
require(tidyr)
require(plotrix)
```

[Use `install.packages("name of package")` if the packages are not yet installed on your system. On some systems the xlsx package might cause problems, so pay attention to the output in the console, in case the packages were not successfully installed. You will still be able to complete this practical, or you can have a look in the next section for some other alternative packages for loading Excel files.]

4.4 PRACTICAL: Import and clean some data

4.4.1 Downloading and importing data

First create a new `.R` script file in your `scripts` folder or equivalent via the drop-down menu or using `Ctrl+Shift+N`. Naming it something like `01-DataImport.R` will make your project management easier in the long run, but feel free to set up your own file naming system – but try to stick with it!

It is good practice to establish a header system for all your script files, such as the one below. The `#` lines are also a good way of making the code file structure easy to understand. Following the Google's R Style Guide rule "The maximum line length is 80 characters.", a nice little trick is to make these separators 80 hashtags long, which gives you a nice visual reference for when your code gets too wide.

```
################################################################################
## DATA IMPORT AND CLEANUP
################################################################################
## 1.1  Import a .csv file
## 1.2  Download and import an Excel file
## 1.3  Download, unzip and import a .dat file
################################################################################
```

Make sure your working directory is at the top of your project folder using `getwd()`. We are going to use the `read.csv()` function to import the data from the `pop2010.csv` file in the `data` folder. But just to be safe, we will first do a test run, importing only 10 rows, so we can inspect the result before importing the whole table:

```
## 1.1  Import a .csv file
################################################################################
# don't forget to note the source of the data! e.g:
# downloaded manually from https://github.com/majazaloznik/RepResCoreSkillsR/
# blob/master/data/pop2010.csv?raw=true" on 20.6.2016
# MZ got it from:
# http://www.census.gov/data/developers/data-sets/international-database.html

getwd()
```

```
## [1] "C:/Users/sfos0247/Dropbox/XtraWork/R stuff/RepResCoreSkillsR"
```

```r
# test run
population2010 <- read.csv("data/pop2010.csv", nrows=10)
population2010
```

```
##    AGE AREA_KM2                 NAME    POP SEX FIPS time
## 1    0      180                Aruba    660   1   AA 2010
## 2    0      180                Aruba    653   2   AA 2010
## 3    0      443  Antigua and Barbuda    720   1   AC 2010
## 4    0      443  Antigua and Barbuda    688   2   AC 2010
## 5    0    83600 United Arab Emirates  40770   1   AE 2010
## 6    0    83600 United Arab Emirates  38987   2   AE 2010
## 7    0   652230          Afghanistan 534585   1   AF 2010
## 8    0   652230          Afghanistan 516673   2   AF 2010
## 9    0  2381741              Algeria 434735   1   AG 2010
## 10   0  2381741              Algeria 414578   2   AG 2010
```

That looks good, the only thing is, I can tell you that all the data in this file is from 2010, so we do not really need the last column. In order to skip it during import, we can use the `colClasses` argument, and setting the seventh argument to `NULL`

```r
# import full table except for 7th column (year)
population2010 <- read.csv("data/pop2010.csv",
                           colClasses = c("integer",   # age
                                          "integer",   # area
                                          "character", # name
                                          "integer",   # population
                                          "integer",   # sex
                                          "character", # country id (FIPS)
                                          "NULL"))     # year - skip
# check how it looks
head(population2010)
```

```
##   AGE AREA_KM2                 NAME   POP SEX FIPS
## 1   0      180                Aruba   660   1   AA
## 2   0      180                Aruba   653   2   AA
## 3   0      443  Antigua and Barbuda   720   1   AC
## 4   0      443  Antigua and Barbuda   688   2   AC
## 5   0    83600 United Arab Emirates 40770   1   AE
## 6   0    83600 United Arab Emirates 38987   2   AE
```

```r
tail(population2010)
```

```
##       AGE AREA_KM2     NAME POP SEX FIPS
## 69079  95   386847 Zimbabwe 506   0   ZI
## 69080  96   386847 Zimbabwe 340   0   ZI
## 69081  97   386847 Zimbabwe 223   0   ZI
## 69082  98   386847 Zimbabwe 142   0   ZI
## 69083  99   386847 Zimbabwe  88   0   ZI
## 69084 100   386847 Zimbabwe 116   0   ZI
```

Comma separated values (`.csv`) is one of the preferred formats to import data from, but R allows you to import from a variety of other formats, although this can sometimes get a bit more messy. This time we will also first download the file, before importing a table from one of the spreadsheets. This is from https://data.gov.uk/dataset/social_trends, part of the government's open data access initiative and a great resource!

```
## 1.2  Download and import an Excel file
###############################################################################
# url of the .xls file we want (using paste only to keep code under 80 chars:)
data.url <- paste("http://www.ons.gov.uk/ons/rel/social-trends-rd/social-",
            "trends/social-trends-41/income-and-wealth-data.xls", sep="")

# download location
data.location <-  paste( "data", "income-and-wealth-data.xls", sep = "/")

# download - Excel files are binary, so set the mode to "wb"!!
download.file(data.url, data.location, mode="wb")
```

You can now have a look in the data folder to check the file has been correctly downloaded and inspect it in Excel. We will import the table from the third worksheet, named "Table 1", on people's perceptions of the current economic situation. Close the Excel file before proceeding! Several solutions are available for importing Excel files into R, a nice overview can be found http://www.r-bloggers.com/read-excel-files-from-r/. In this practical we will use the `xlsx` package:

```
## Importing the data from an .xls file
require(xlsx)

# let's see what happens if we import the whole sheet
economic.situation <- read.xlsx(data.location, sheetIndex = 3)
```

Have a look at `economic.situation`. [2] It is not ideal, empty rows and columns are imported, as is the text at the top and the bottom of the worksheet. Luckily, `read.xlsx` has plenty of arguments that allow us to specify more precisely what we want to import. In this case, we can go one step further, and note that there are actually three separate tables in this worksheet, so it might be easiest to import them separately:

```
## using rowIndex and colIndex select each subtable individually:
world.situation <- read.xlsx(data.location, sheetIndex = 3,
                         rowIndex=c(4, 6:8), colIndex = c(1:7))

UK.situation <- read.xlsx(data.location, sheetIndex = 3,
                           rowIndex=c(4, 11:13), colIndex = c(1:7))

household.situation <- read.xlsx(data.location, sheetIndex = 3,
                          rowIndex=c(4, 16:18), colIndex = c(1:7))
```

Finally, sometimes we need to extract the data from a zipped file, this can also be done directly form R [3]. And to try out another format we will import an SPSS file as well.[4]

```
## 1.3  Download, unzip and import a .dat file
###############################################################################
# url of the .zip file we want
data.zip.url <- paste("http://spss.allenandunwin.com.s3-website-ap-southeast-2.",
                  "amazonaws.com/Files/survey.zip", sep="")
temp <- tempfile()
download.file(data.zip.url, temp)
# check what is in the zip file using list (doesn't extract anything)
unzip(temp, list=TRUE)
```

```
##         Name Length                Date
```

[2] Are you getting an error? That may be because you still have the Excel file open!

[3] This should work even if no winzip utility is installed on the machine?

[4] The data file is supplementary material to the SPSS Survival Manual from a survey designed to explore the factors that impact on respondents' psychological adjustment and well-being.

```
## 1 survey.sav  84640 2015-12-22 12:09:00
```

```
# Only one file, that's the one we want to extract to the data folder
unzip(temp, "survey.sav", exdir = "data")
unlink(temp)
```

You can now check the `data` folder and you should find the `survery.sav` file there. Even if you don't have SPSS installed on your computer, you can now open it using R and the `foreign` package:

```
require(foreign)

# import the data as a data frame:
data.location <- paste("data","survey.sav", sep="/")
ed.psy.survey <- read.spss(data.location,  to.data.frame=TRUE)
# check what it looks like
ed.psy.survey[1:5,1:5]
```

```
##    id     sex age             marital child
## 1 415 FEMALES  24  MARRIED FIRST TIME   YES
## 2   9   MALES  39 LIVING WITH PARTNER   YES
## 3 425 FEMALES  48  MARRIED FIRST TIME   YES
## 4 307   MALES  41           REMARRIED   YES
## 5 440   MALES  23              SINGLE    NO
```

Now we have the data, before we continue we'll just do a bit of housekeeping and clear our workspace of the objects we don't need any more (including the survey dataset, which we will not use in this practical)

```
# CLEAN UP!
rm(economic.situation, ed.psy.survey, data.location, data.url, data.zip.url, temp)
```

4.4.2 Data Tidying

In the second part of this practical we will use the functions from the `tidyr` package to tidy up the two datasets.

```
## 2.1 tidy up the population data
###############################################################################
require(tidyr)
```

The population2010 data.frame is already pretty tidy! The only issue with it is the `SEX` variable, which is coded for men (`SEX==1`), women (`SEX==2`), and both (`SEX==0`). We really only need to remove the rows with the values for (`SEX==0`), but we can use this opportunity to perform a data check as well, while practising the `spread()` and `gather()` functions:

First let's try out our technique on a small subset of the data - this is good practice in general, especially if you are dealing with large datasets. We'll select only the observations for Aruba, and have a look at them:

```
# try out our technique on a smaller subset of the data
test.data <- population2010[population2010$FIPS == "AA", ]
head(test.data)
```

```
##     AGE AREA_KM2  NAME POP SEX FIPS
## 1     0      180 Aruba 660   1   AA
## 2     0      180 Aruba 653   2   AA
## 457   1      180 Aruba 651   1   AA
## 458   1      180 Aruba 645   2   AA
## 913   2      180 Aruba 643   1   AA
## 914   2      180 Aruba 636   2   AA
```

We want to reshape the table so that the values of `SEX` will become new column names (i.e. *keys*), and that the *values* for these new keys will be the values from the variable `POP`. This means the `spread()` functions should look like this:

```
tidy.test <- spread(test.data, SEX, POP )
head(tidy.test)
```

```
##   AGE AREA_KM2  NAME FIPS    0   1   2
## 1   0      180 Aruba   AA 1313 660 653
## 2   1      180 Aruba   AA 1296 651 645
## 3   2      180 Aruba   AA 1279 643 636
## 4   3      180 Aruba   AA 1266 634 632
## 5   4      180 Aruba   AA 1251 627 624
## 6   5      180 Aruba   AA 1243 624 619
```

```
# and for clarity, let's rename the columns:
colnames(tidy.test)[5:7] <- c("both", "male", "female")
```

We can now check if the totals for men and women actually match, before we discard the column with the sum of both:

```
# calculate sum of males and females
tidy.test$check <- tidy.test$male + tidy.test$female

# compare it with the values already in the table:
all.equal(tidy.test$both, tidy.test$check)
```

```
## [1] TRUE
```

```
# looks good, now we can remove both total columns:
tidy.test$check <- NULL
tidy.test$both <- NULL

# so now we have:
head(tidy.test)
```

```
##   AGE AREA_KM2  NAME FIPS male female
## 1   0      180 Aruba   AA  660    653
## 2   1      180 Aruba   AA  651    645
## 3   2      180 Aruba   AA  643    636
## 4   3      180 Aruba   AA  634    632
## 5   4      180 Aruba   AA  627    624
## 6   5      180 Aruba   AA  624    619
```

And finally we have to use `gather()` to get back to a tidy table. Remember, with gather you need to pass the *names* of the new variables that are now the *key* and the *value*, and the column names which hold them:

```
tidy.test <- gather(tidy.test, sex, population, 5:6)
# and let's check it again:
head(tidy.test)
```

```
##   AGE AREA_KM2  NAME FIPS  sex population
## 1   0      180 Aruba   AA male        660
## 2   1      180 Aruba   AA male        651
## 3   2      180 Aruba   AA male        643
## 4   3      180 Aruba   AA male        634
## 5   4      180 Aruba   AA male        627
## 6   5      180 Aruba   AA male        624
```

If you are happy with the test run, you can now try it on the whole `population2010` table.

```
head(tidy.population2010)
```

```
##   AGE AREA_KM2             NAME FIPS  sex population
## 1   0        2           Monaco   MN male        110
## 2   0        7        Gibraltar   GI male        209
## 3   0       21            Nauru   NR male        115
## 4   0       21 Saint Barthelemy   TB male         41
## 5   0       26           Tuvalu   TV male        119
## 6   0       28            Macau   MC male       2586
```

```
## 2.2 tidy up the perception data
###############################################################################
```

Most of the time you will not be lucky enough to work with as nicely formed datasets as the population one. But the same tools can be used to disentangle much more messy tables, such as the ones we extracted from the Excel file above.

Let's have a look at one of the three files, e.g. `household.situation`, and see how it could be tidied up. What are the variables (that should be in the columns), and what are the observations (that should have one row each)?

```
# first let's remane the column names
colnames(household.situation) <- c("perception", "inc.lt.20",
                                   "inc.20.to.39", "inc.40.to.59",
                                   "inc.60.to.99", "inc.gt.100", "all")
household.situation
```

```
##               perception inc.lt.20 inc.20.to.39 inc.40.to.59 inc.60.to.99
## 1      Good or very good        28           44           54           64
## 2    Neither good or bad        47           42           35           36
## 3        Bad or very bad        25           14           11            1
##   inc.gt.100 all
## 1         63  40
## 2         34  43
## 3          3  18
```

In fact the whole table needs to be transposed, so that each population group represents one observation, and the proportion answering each question are the variables. In order to do that we need to first gather the data in long form, before spreading it out again wide.[5]

```
# transpose using gather and spread
X.household.situation <- gather(household.situation, income.group, proportion, 2:7)
tidy.household.situation <- spread(X.household.situation, perception, proportion)
# let's also rename the column names in keeping with the convention of avoiding spaces
colnames(tidy.household.situation) <- c("income.group", "bad", "good", "neutral")
# check the result and remove the temporary table
tidy.household.situation
```

```
##   income.group bad good neutral
## 1    inc.lt.20  25   28      47
## 2 inc.20.to.39  14   44      42
## 3 inc.40.to.59  11   54      35
## 4 inc.60.to.99   1   64      36
## 5   inc.gt.100   3   63      34
## 6          all  18   40      43
```

[5] The `t()` function will transpose a data frame in R, try it out to see if it is a useful alternative to gather and spread.

```
rm(X.household.situation)
```

Don't forget, we have two more tables, one for each perception question. If we want to merge them together at the end, we need to be clear that each observation refers to one of the questions:

```
tidy.household.situation$perception <- "HH"
```

Now you can repeat the tidying for the UK and world perceptions, and once all three tables are tidy, you can merge them together using `rbind()`:

```
# merge all the tables together
tidy.economic.situation <- rbind(tidy.household.situation,
                                 tidy.UK.situation,
                                 tidy.world.situation)
```

Finally, you can now clear your workspace using `rm()` as we did before, to remove everything except for `tidy.population2010` and `tidy.economic.situation`.

5 Efficient Coding

This section covers some of the most important skills to improve the efficiency, readability, and reproducibility of your R code. The standard control of the flow of your code that can be achieved with `ifelse` statements and looping is covered briefly, however you are encouraged in particular to explore the advantages of *vectorised* R code in particular the `apply` family of functions. Writing your own functions will greatly streamline your work, as well as forcing you to think in more abstract terms about your analysis - making it easily transferable and reproducible as opposed to limited to the specific situation you find yourself analysing at the moment.

5.1 Standard control structures

An indispensable gain in efficiency of your programming can be achieved by using *control strucures* to control the execution of your code. These can be divided into *conditional execution* structures (`if` and `else` type functions) and *looping* structures. However, as we shall see in the next section, there are some some pretty good ways (and reasons) to avoid looping in R.

5.1.1 Conditional execution

The standard syntax for conditional execution is as follows:

```
if (condition) {
  # do something
} else {
  # do something else
}
```

In fact, you may also use only the `if()` construct on it's own:

```
if (condition) {
  # do something
}
```

The if/else syntax also works in a single line, where you can dispense with the curly braces:

```
if (x >= 0)  print("Poz") else print("Neg")
```

While this is more compact, it can impact readability, and can also make your code more difficult to debug and extend. Using curly braces and indenting the code properly will make it clearer to the reader, and also easier to e.g. extend via nesting:

```
x <- runif(1) # randum number from uniform distribution [0,1]
if (x >= 0.6) {
  print("Good")
} else {
  if (x <= 0.4) {
    print("Bad")
  } else {
    print("Not Sure")}
}
```

The conditions to be evaluated are:

```
x == y   # x is equal to y
x != y   # x is not equal to y
x > y    # x is greater than y
x < y    # x is less than y
```

```
x <= y   # x is less than or equal to y
x >= y   # x is greater than or equal to y
x %in% y # x is located in y
TRUE     #
FALSE    #
```

And these can further be combined using standard logical operators:

```
! x       # NOT
x & y     # AND
x | y     # OR
xor(x, y) # exclusive OR
```

What if you want to run a conditional statement over an entire vector? You might be tempted to jump to the next section on looping, and construct a loop going over each element of the vector and evaluating the condition. This would of course work, but it would be a very inefficient way of coding, and would not be taking advantage of the efficiencies of vectorisation in R: vectorised functions apply to whole vectors at once instead of evaluating for each element individually. We should therefore use the *vectorised* form of the if/else construct `ifelse()`:

```
ifelse(condition, yes, no)
```

Here `yes` is the value to be returned if the condition is satisfied, and `no` if not. Similarly as above, `ifelse()` statements can also be nested as in this example:

```
x <- runif(20) # 20 randum numbers from uniform distribution [0,1]
ifelse(x >= 0.6, "G",
       ifelse(x<=0.4, "B", "N"))
```

```
##  [1] "G" "N" "N" "N" "G" "B" "N" "G" "B" "G" "B" "B" "B" "B" "B" "N" "G"
## [18] "G" "B" "N"
```

5.1.2 Looping

R distinguishes two types of loops:

- ones that execute a function a predetermined number of times, as determined by an *index* [`i`]
- ones that execute a function until a condition is met

The `for()` loop construct takes the following form:

```
for (i in seq) expr
```

Again, using curly braces is usually preferred, for loops can be nested and the indices need not be integers:

```
mat <- matrix(NA, nrow=3, ncol=3)
for (i in 1:3){
  for (j in 1:3){
    mat[i,j] <- paste(i, j, sep="-")
  }
}
mat
```

```
##      [,1]  [,2]  [,3]
## [1,] "1-1" "1-2" "1-3"
## [2,] "2-1" "2-2" "2-3"
## [3,] "3-1" "3-2" "3-3"
```

While loops take the following form:

```
while(cond) expr
```

```
cumsum <- 0
while(cumsum <= 3) {
  cumsum <- cumsum + runif(1)
  print(cumsum)
}
```

```
## [1] 0.5876
## [1] 1.34
## [1] 1.923
## [1] 2.776
## [1] 3.421
```

A repeat loop is similar, but we must explicitly add a `break` to specify when to exit the loop:

```
cumsum <- 0
repeat {
  cumsum <- cumsum + runif(1)
  print(cumsum)
  if (cumsum >= 3) break
}
```

```
## [1] 0.8262
## [1] 1.09
## [1] 1.203
## [1] 2.16
## [1] 2.498
## [1] 2.772
## [1] 3.613
```

Both of these constructs should be used with great care, as careless specification of the exiting condition can leave you stuck in an infinite loop. Try running the last example without the line specifying the break! Luckily RStudio allows you to interrupt such an endless loop using the little red stop button in the top right corner of the console window.

5.2 Vecotrisation and `apply` family of funcitons

Looping functions - the `for()` loop in particular - are very intuitive and mastering them can represent a quick capability boost for a new R programmer. It is however highly recommended that you spend some time mastering the related `apply` family of functions, which should cover most of your looping needs. The genera rule is this: If you need to apply an expression over a series of elements and the *order* in which you do this is important, then use a loop. If the order is not important, take advantage of `apply`. In many circumstances this can improve the speed of your code, but in all cases it will make your code simpler and easier to read.

The underlying logic of the `apply` family is starting out with some data structure (a vector, matrix, data.frame etc.), we want to split it into constituent parts, apply a function on each of them, and combine them back[6]. We might for example want to apply a function on every row of a data.frame, every element of a vector, or every column in a matrix.

[6]This idea comes from Hadley Wickham's paper on the split-apply-combine strategy of data analysis: http://vita.had.co.nz/papers/plyr.html

5.2.1 apply()

The `apply()` function will apply a function to either the rows or the columns of a matrix. It's basic structure is:

```
apply(X, MARGIN, FUN, ...)
```

Where X is a matrix (if it is a data.frame, R will coerce it to a matrix), `MARGIN == 1` indicates rows, and `MARGIN == 2` indicates columns. A simple example of its use is to calculate row and column totals:

```
mat <- matrix(1:9, 3,3)
# row totals
apply(mat, 1, sum)
```

```
## [1] 12 15 18
```

```
# column totals
apply(mat, 2, sum)
```

```
## [1]  6 15 24
```

In passing the function `FUN` in the example here we used built in function `sum`, but the real power of `apply` comes from integrating it with user defined functions. These are covered in the next section, but here is a quick example of how an in-line function can be used to find the second largest value in each row of a matrix.

```
mat <- matrix(sample(1:100, 25), 5,5)
mat
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]   32   74   22   40    1
## [2,]   13   98   60   31   43
## [3,]   41   42   96   86   70
## [4,]   97   57   68   21    3
## [5,]   39   25   76   72   80
```

```
# find the second largest value in each row
apply(mat, 1, function(x) sort(x, decreasing = TRUE)[2])
```

```
## [1] 40 60 86 68 76
```

```
# and for comparison, here is how we would do this using a for loop
out <- vector()
for (i in 1:nrow(mat)) {
  out[i] <- sort(mat[i,], decreasing = TRUE)[2]
}
out
```

```
## [1] 40 60 86 68 76
```

5.2.2 lapply() and sapply()

The functions `lapply()` and `sapply()` both apply a function to a vector, and the first returns a list back, while the second will try to simplify and return a vector.

It is important to note that in R there are two types of vectors: i) atomic vectors and ii) lists.

Furthermore, data frames in R are also represented as lists, with each column is an element of the list, represented by a vector.

So this means both these functions can be applied to atomic vectors, to data frames, or to other types of lists:

```r
# a list of elements with different lengths:
test <- list(a = 1:5, b = 20:100, c = 17234)
lapply(test, min)
```

```
## $a
## [1] 1
##
## $b
## [1] 20
##
## $c
## [1] 17234
```

```r
sapply(test, min)
```

```
##     a     b     c
##     1    20 17234
```

```r
# a data frame (list of three vectors of equal length):
test <- data.frame(a = 1:5, b = 6:10, c = 11:15)
lapply(test, mean)
```

```
## $a
## [1] 3
##
## $b
## [1] 8
##
## $c
## [1] 13
```

```r
sapply(test, mean)
```

```
##  a  b  c
##  3  8 13
```

```r
# an atomic vector (this is rather silly, since sqrt(X) would work the same)
# but is added for completeness
test <- 1:3
lapply(test, sqrt)
```

```
## [[1]]
## [1] 1
##
## [[2]]
## [1] 1.414
##
## [[3]]
## [1] 1.732
```

```r
sapply(test, sqrt)
```

```
## [1] 1.000 1.414 1.732
```

By writing more elaborate functions and passing them as the argument to any of the `apply` family of functions, this seemingly simple construct can become incredibly powerful - as well as making the code eminently readable.

5.3 Writing your own functions

One of the greatest strengths of R comes from writing your own functions. This not only allows you to repeat the same procedure consistently, but makes your code more structured and readable reduces chance of error, and will further strengthens your reproducibility mentality.

The basic construct is as follows:

```
function.name <- function(arguments, ...) {
  expression
  (return value)
}
```

we have already seen the in-line version of the function call in the apply example above, remember:

```
function(x) sort(x, decreasing = TRUE)[2]
```

Here our function takes a single argument (`x`), evaluates the expression (`sort(x, decreasing = TRUE)[2]`), and returns the value of that expression. This only works if the function has only a single expression, in which case the evaluated expression is returned. Otherwise we have to explicitly state what we want returned. We can rewrite this function in the more elaborate mode:

```
FunSecondLargest <-function(x) {
  r <- sort(x, decreasing = TRUE)[2]
  return(r)
}
# now let's try it out with a sample vector
test.vector <- tidy.population2010$population

FunSecondLargest(test.vector)
```

```
## [1] 14642884
```

We can also now use this function directly in the apply call we used before:

```
apply(mat, 1, FunSecondLargest)
```

```
## [1] 40 60 86 68 76
```

We can also quickly rewrite the function to instead find the n-th largest value in the vector, by adding an additional argument `n`. And don't forget to write sensible commentary about what you are doing - at least for the benefit of your future self!

```
# Function for extracting the n-th largest value from a vector
# Agruments:
#   x - vector
#   n - optional integer value for rank
# Output:
# Returns single value
FunNthLargest <-function(x, n=1) {
  r <- sort(x, decreasing = TRUE)[n]
  return(r)
}
# by default n=1, so it will find the largest value if we don't specify
FunNthLargest(test.vector)
```

```
## [1] 15120232
```

```
FunNthLargest(test.vector, n=2)
```

```
## [1] 14642884
```

```r
FunNthLargest(test.vector, n=3)
```

```
## [1] 13601669
```

We could e.g. further generalise this function to look for the n-th smallest value, by adding another argument for the `TRUE/FALSE` value that gets passed to `decreasing` etc. Note that unlike the argument `x`, the argument `n` has a default value (=1). This means we do not have to explicitly specify it unless we want it to be a different value.

Functions have their own local environment, which is not accessible from the global environment. This means that whatever calculations are evaluated inside the function call do not clutter your workspace, but also their results are not accessible unles you explicitly return them from the function. Thus the object `r` will not be found in the global environment, instead we will get the error:

```r
r
Error: object 'r' not found
```

We can also have our function return several outputs for example:

```r
FunNthLargestElaborate <-function(x, n=1) {
  r <- sort(x, decreasing = TRUE)[n]
  desc <- paste("Rank", n, sep=":")
  return(c(desc, r))
}
FunNthLargestElaborate(test.vector, 3)
```

```
## [1] "Rank:3"   "13601669"
```

As we have seen before with if/else statements and loops, functions can also be nested – as well as combined with if/else statements and loops! It is good practice to try to keep your code modular: keep your functions short and call them from each other. This again makes it easier for the reader to understand what is going on, and easier for you to find errors or update your code.

From a project management point of view it is also good practice to store all your functions in a separate file, which you `source()` at the beginning of each session. You can even add `source("00-MyFunctions.R")` to your `.RProfile` file, which means all your bespoke functions will be automatically uploaded at the start of each session.

5.4 PRACTICAL: If/else, loops, apply and functions

```
## 3.1 Practice conditional expressions and logical operators
###############################################################################
```

Practice conditional expressions and logical operators by seeing if you can figure out the results of the following expressions, then check them in R:

```
x <- 1:7
y <- -3:3
x
```

```
## [1] 1 2 3 4 5 6 7
```

```
y
```

```
## [1] -3 -2 -1  0  1  2  3
```

```
# try the following:
(x == y)
(x > abs(y))
(x > 3) & (x < 5)
(x > 3) | (x < 5)
xor((x > 3), (x < 5))
(-1 %in% y)
(3 %in% y) & (3 %in% x)
(3 %in% y) & !(3 %in% x)
```

```
## 3.2 Check consistency of tidy.economic.situation proportions using IF - ELSE
###############################################################################
```

Use a `for()` loop to go through every row of `tidy.economic.situation` (tip: `nrow()` will tell you how many iterations you need):

- for each row add up the proportions for all three answers (columns two to four)
- use an `if/else` construct to check if the total equals 100
  - if it does, use `print` to print out an OK message
  - if it doesn't, print out a different message, one created using `paste` - so you can include the information on *which* row you have found the error.

Use the following template to write your loop:

```
for (i in 1: nrow(tidy.economic.situation)) {
  if(???){
    ???} else {
      ???}
}
```

```
## 3.3 Write a function to check consistency of tidy.economic.situation proportions
###############################################################################
```

Now write a function for the row checking you just did inside the `for()` loop. This is simply generalising the if/else expression to take a supplied argument instead of explicitly naming the row:

- Make sure you *document* your function correctly!
- the input for the function should be a row
- the output of the function should be a variable called `message` - "OK" or "Not OK"
- Use the framework below:

```r
FunRowCheck <- function(x) {
  message <- if(???){
    ???} else {
      ???}
  return(message)
}
```

Now test out your function on a single row:

```r
FunRowCheck(tidy.economic.situation[1,2:4])
```

```
## [1] "OK"
```

If it works correctly, you can now try using your new function inside an `apply` construct.

Remember, `apply` will evaluate the expression along the *whole* row, and you want to apply it only to columns 2:4, so make sure you don't pass the whole table to `apply`. When you are happy with the result, append it to the table as an additional column:

```r
tidy.economic.situation$test <- apply(???)
```

Your table should now look like this:

```r
tidy.economic.situation
```

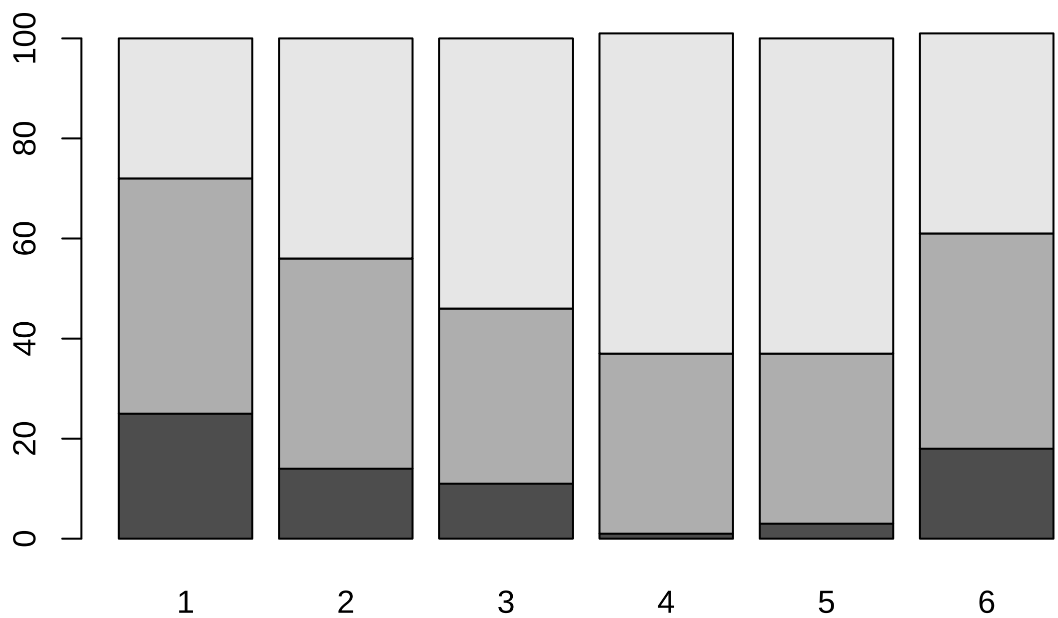
```
##    income.group bad good neutral perception   test
## 1     inc.lt.20  25   28      47         HH     OK
## 2  inc.20.to.39  14   44      42         HH     OK
## 3  inc.40.to.59  11   54      35         HH     OK
## 4  inc.60.to.99   1   64      36         HH Not OK
## 5    inc.gt.100   3   63      34         HH     OK
## 6           all  18   40      43         HH Not OK
## 7     inc.lt.20  76    8      17         UK Not OK
## 8  inc.20.to.39  80    5      15         UK     OK
## 9  inc.40.to.59  83    4      13         UK     OK
## 10 inc.60.to.99  92    4       4         UK     OK
## 11   inc.gt.100  89    0      11         UK     OK
## 12          all  80    6      15         UK Not OK
## 13    inc.lt.20  77    6      17          W     OK
## 14 inc.20.to.39  79    3      18          W     OK
## 15 inc.40.to.59  85    2      12          W Not OK
## 16 inc.60.to.99  91    1       8          W     OK
## 17   inc.gt.100  92    0       8          W     OK
## 18          all  80    4      16          W     OK
```

If it doesn't have a look at the `02-FunctionsAndLoops.R` file in the scripts folder for the solution before you proceed to the next task.

```r
## 3.5 Bar-Plotting function
###############################################################################
```

To finish off this practical session we will write a function for plotting our table, using the `barplot()` function. This function only accepts vectors or matrices, but because our data is in a data.frame, we need to use `as.matrix()` for it to work, in addition to `t()` for transposing it. Here is the code for the most stripped down stacked barplot of the people's perceptions of their household financial situation. Note that the order of the columns had to be changed to make the more logical order of bad - neutral - good.

```r
barplot(t(as.matrix(tidy.economic.situation[1:6,c(2,4,3)])))
```



Now expand the barplot function - use the help documentation in the help tab:

- add names to the x-axis
- add legend text
- add a main title
- feel free to explore additional arguments to the plot!

Once you are happy with your plot, enclose it in a function, so that you can pass it each of the three subsets of the table individually, and the function will additionally also change the plot's title to the correct one.

6 Data manipulation with dplyr

Finally, the relatively new `dplyr` family of functions is one of the most powerful recent developments in the R coding world. While all of the data processing capabilities of `dplyr` existed in one shape or another in R before (usually several), `dplyr` brings them together in a comprehensive and systematic way, that allows for cleaner code that is easier to read, and faster to run. It also integrates logically with the `tidyr` family of functions described earlier, but it's most exciting and revolutionary aspect is it's assimilation of the *piping* or *chaining* of successive data processing functions - originally developed in the `magrittr` package, and now rightly becoming mainstream R practice.

We will cover some of the most important functionalities of `dplyr` here, but you are encouraged to explore several excellent on-line resources, including video tutorials etc. Keep the same Data Wrangling Cheatsheet that also covers `tidyr` to hand - for your convenience a copy is also stored in the `literature` folder for this course's repository.

For this whole section we will be using the two data tables prepared in the first practical.

6.1 Subsetting

6.1.1 filter()

Extracts rows that meet the logical criteria:

```
filter(data, criteria)
```

You can use any of the evaluation conditions and logical operators we covered at the beginning of the previous section:

```
filter(tidy.population2010, AREA_KM2 < 1000 & population > 10000 & AGE == 0)
```

```
##   AGE AREA_KM2       NAME FIPS    sex population
## 1   0      360 Gaza Strip   GZ   male      29209
## 2   0      687  Singapore   SN   male      18632
## 3   0      360 Gaza Strip   GZ female      27616
## 4   0      687  Singapore   SN female      17438
```

6.1.2 select()

Extracts only the columns that you list:

```
select(data, list)
```

```
select(tidy.economic.situation, bad, good, neutral)
```

```
##    bad good neutral
## 1   25   28      47
## 2   14   44      42
## 3   11   54      35
## 4    1   64      36
## 5    3   63      34
## 6   18   40      43
## 7   76    8      17
## 8   80    5      15
## 9   83    4      13
## 10  92    4       4
## 11  89    0      11
```

```
## 12  80    6       15
## 13  77    6       17
## 14  79    3       18
## 15  85    2       12
## 16  91    1        8
## 17  92    0        8
## 18  80    4       16
```

In addition there is a large number of helper functions to select column names:

```r
# all the columns between bad and neutral
head(select(tidy.economic.situation, bad:neutral), n=3)
```

```
##   bad good neutral
## 1  25   28      47
## 2  14   44      42
## 3  11   54      35
```

```r
# all but the perception column
head(select(tidy.economic.situation, -perception), n=3)
```

```
##   income.group bad good neutral test
## 1    inc.lt.20  25   28      47   OK
## 2 inc.20.to.39  14   44      42   OK
## 3 inc.40.to.59  11   54      35   OK
```

```r
# contains a dot in the name:
head(select(tidy.economic.situation, contains(".")), n=3)
```

```
##   income.group
## 1    inc.lt.20
## 2 inc.20.to.39
## 3 inc.40.to.59
```

```r
# starts with the letter p
head(select(tidy.economic.situation, starts_with("p")), n=3)
```

```
##   perception
## 1         HH
## 2         HH
## 3         HH
```

```r
# ends with the letter d
head(select(tidy.economic.situation, ends_with("d")), n=3)
```

```
##   bad good
## 1  25   28
## 2  14   44
## 3  11   54
```

```r
# contains the text "n"
head(select(tidy.economic.situation, contains("n")), n=3)
```

```
##   income.group neutral perception
## 1    inc.lt.20      47         HH
## 2 inc.20.to.39      42         HH
## 3 inc.40.to.59      35         HH
```

You can also use select to reorder the columns, in our case we might want to reorder the three answer columns so:

```
# change order of columns
head(select(tidy.economic.situation,income.group:bad, neutral, good:perception), n=3)
```

```
##   income.group bad neutral good perception
## 1    inc.lt.20  25      47   28         HH
## 2 inc.20.to.39  14      42   44         HH
## 3 inc.40.to.59  11      35   54         HH
```

```
# we can also do this using the columns' respective indices instead
head(select(tidy.economic.situation, 1,2,4,3,5), n=3)
```

```
##   income.group bad neutral good perception
## 1    inc.lt.20  25      47   28         HH
## 2 inc.20.to.39  14      42   44         HH
## 3 inc.40.to.59  11      35   54         HH
```

6.2 Making new variables

New variables are easily created using `mutate()`, which has the additional advantage of allowing you to reuse variables as you create them, without the need for an intermediate step!

```
# scale the values so they all sum up to 100
tidy.economic.situation <- mutate(tidy.economic.situation, total = bad+neutral+good,
                                  bad.scaled = bad/total*100,
                                  good.scaled = good/total*100,
                                  neutral.scaled = neutral/total*100,
                                  total.scaled =  bad.scaled  + good.scaled +
                                    neutral.scaled
)
head(tidy.economic.situation)
```

```
##   income.group bad good neutral perception   test total bad.scaled
## 1    inc.lt.20  25   28      47         HH     OK   100    25.0000
## 2 inc.20.to.39  14   44      42         HH     OK   100    14.0000
## 3 inc.40.to.59  11   54      35         HH     OK   100    11.0000
## 4 inc.60.to.99   1   64      36         HH Not OK   101     0.9901
## 5   inc.gt.100   3   63      34         HH     OK   100     3.0000
## 6          all  18   40      43         HH Not OK   101    17.8218
##   good.scaled neutral.scaled total.scaled
## 1       28.00          47.00          100
## 2       44.00          42.00          100
## 3       54.00          35.00          100
## 4       63.37          35.64          100
## 5       63.00          34.00          100
## 6       39.60          42.57          100
```

```
# we can now use select to remove the old ones
tidy.economic.situation <- select(tidy.economic.situation, -bad, - good, -neutral,
                                  -total, -total.scaled)
# we could also use rename to rename the new ones
tidy.economic.situation <- rename(tidy.economic.situation, bad = bad.scaled,
                                  good = good.scaled, neutral = neutral.scaled)
```

New variables can also be made using existing or user written functions. For example using `cut()` we can recode the bad variable into a categorical one:

```r
head(mutate(tidy.economic.situation, bad.cat = cut(bad, seq(0,100,10))))
```

```
##   income.group perception   test     bad  good neutral bad.cat
## 1    inc.lt.20         HH     OK 25.0000 28.00   47.00 (20,30]
## 2 inc.20.to.39         HH     OK 14.0000 44.00   42.00 (10,20]
## 3 inc.40.to.59         HH     OK 11.0000 54.00   35.00 (10,20]
## 4 inc.60.to.99         HH Not OK  0.9901 63.37   35.64  (0,10]
## 5   inc.gt.100         HH     OK  3.0000 63.00   34.00  (0,10]
## 6          all         HH Not OK 17.8218 39.60   42.57 (10,20]
```

6.3 Summarizing

We can quickly summarise the data column wise using the `summarise()` function

```r
summarise(data, new.var = summary.function(column))
```

For example the average population, average area and total count in the population table, we can also use our own functions, such as the one we wrote before to get the second largest population value

```r
summarise(tidy.population2010, pop = mean(population), area = mean(AREA_KM2), count = n(),
          test = FunSecondLargest(population))
```

```
##      pop   area count     test
## 1 149087 578149 46056 14642884
```

But the summarise function really comes into it's own when it operates on a *grouped* table. Using the function `group_by()` the table is (invisibly) split into sub-tables by the values of the grouping variable, and the summarise function then operates on each subset individually:

```r
summarise(group_by(tidy.population2010, AGE),
          av.pop = mean(population), av.area = mean(AREA_KM2), count = n(),
          test = FunSecondLargest(population))
```

```
## Source: local data frame [101 x 5]
##
##      AGE av.pop av.area count     test
##    (int)  (dbl)   (dbl) (int)    (int)
## 1      0 282738  578149   456 11355900
## 2      1 278558  578149   456 11177984
## 3      2 275981  578149   456 11082923
## 4      3 272960  578149   456 11021599
## 5      4 270025  578149   456 11002862
## 6      5 267803  578149   456 11022271
## 7      6 266032  578149   456 11037275
## 8      7 264164  578149   456 11040174
## 9      8 262954  578149   456 11030910
## 10     9 262510  578149   456 11016559
## ..   ...    ...     ...   ...      ...
```

An important corollary to the grouping function is `ungroup()`, which removes the grouping from the table for further analysis – we will use it in the last section of this chapter.

6.4 Joining tables

The dplyr package also contains a set of functions that allow you to join tables by matching on common variables namely:

- left_join(a, b)
- right_join(a, b) – keeps all
- inner_join(a, b) – only keeps rows present in both a and b
- full_join(a, b) – keeps all rows

```
# prepare two small tables, one of UK men aged 0 or 1, the second of women aged 1 or 2:
UK.men <- filter(tidy.population2010, FIPS == "UK" & sex == "male" & (AGE == 0 | AGE == 1 ))
UK.women <- filter(tidy.population2010, FIPS == "UK", sex == "female" & (AGE == 1 | AGE == 2 ))

# try out all 4 merges on the two tables
left_join(UK.men, UK.women, by = c("AGE", "NAME", "AREA_KM2", "FIPS"))
```

```
##   AGE AREA_KM2           NAME FIPS sex.x population.x  sex.y population.y
## 1   0   241930 United Kingdom   UK  male       392514   <NA>           NA
## 2   1   241930 United Kingdom   UK  male       392859 female       373769
```

```
right_join(UK.men, UK.women, by = c("AGE", "NAME", "AREA_KM2", "FIPS"))
```

```
##   AGE AREA_KM2           NAME FIPS sex.x population.x  sex.y population.y
## 1   1   241930 United Kingdom   UK  male       392859 female       373769
## 2   2   241930 United Kingdom   UK  <NA>           NA female       374206
```

```
inner_join(UK.men, UK.women, by = c("AGE", "NAME", "AREA_KM2", "FIPS"))
```

```
##   AGE AREA_KM2           NAME FIPS sex.x population.x  sex.y population.y
## 1   1   241930 United Kingdom   UK  male       392859 female       373769
```

```
full_join(UK.men, UK.women, by = c("AGE", "NAME", "AREA_KM2", "FIPS"))
```

```
##   AGE AREA_KM2           NAME FIPS sex.x population.x  sex.y population.y
## 1   0   241930 United Kingdom   UK  male       392514   <NA>           NA
## 2   1   241930 United Kingdom   UK  male       392859 female       373769
## 3   2   241930 United Kingdom   UK  <NA>           NA female       374206
```

All of these have a `by=` argument, which lets you choose the columns to be joined by - if you do not explicitly name them, `dplyr` uses all the ones with identical names in both tables. If the columns you want to join by have different tables in each table you can specify this so: `by = c("name.a" = "name.b")`

6.4.1 Other `dplyr` funcitons

The Data Wrangling Cheatsheet is an indispensable help with `dplyr` functions. We will briefly mention only one more, but there are several more that we will not cover here.

`arrange()` for data sorting - ascending by default, otherwise specify `desc()`:

```
arrange(tidy.economic.situation, perception, desc(bad))
```

```
##    income.group perception   test     bad   good neutral
## 1     inc.lt.20         HH     OK 25.0000 28.000   47.00
## 2           all         HH Not OK 17.8218 39.604   42.57
## 3  inc.20.to.39         HH     OK 14.0000 44.000   42.00
## 4  inc.40.to.59         HH     OK 11.0000 54.000   35.00
## 5    inc.gt.100         HH     OK  3.0000 63.000   34.00
```

```
## 6  inc.60.to.99         HH Not OK  0.9901 63.366   35.64
## 7  inc.60.to.99         UK     OK 92.0000  4.000    4.00
## 8    inc.gt.100         UK     OK 89.0000  0.000   11.00
## 9  inc.40.to.59         UK     OK 83.0000  4.000   13.00
## 10 inc.20.to.39         UK     OK 80.0000  5.000   15.00
## 11          all         UK Not OK 79.2079  5.941   14.85
## 12     inc.lt.20        UK Not OK 75.2475  7.921   16.83
## 13    inc.gt.100         W     OK 92.0000  0.000    8.00
## 14 inc.60.to.99          W     OK 91.0000  1.000    8.00
## 15 inc.40.to.59          W Not OK 85.8586  2.020   12.12
## 16          all          W     OK 80.0000  4.000   16.00
## 17 inc.20.to.39          W     OK 79.0000  3.000   18.00
## 18     inc.lt.20         W     OK 77.0000  6.000   17.00
```

6.5 Piping/chaining daisies

Piping data brings a completely new level of intuitiveness you R programming. Instead of nesting and indenting successive functions, which means the code has to be read *inside out*, piping (also known as daisy chaining) allows the code to be written in the natural direction in which the data is flowing. The piping operator `%>%` indicates the direction of this flow as well, taking the output of the preceding function and directing it into the next one.

Piping can be applied to almost any function, but shines particularly brightly when combining the dplyr functions we have just covered. For a pretty silly example:

```
tidy.population2010 %>%
  filter(AREA_KM2 < 2000 & population > 15000 & AGE == 0) %>%
  select(-AGE) %>%
  mutate(density = population/AREA_KM2) %>%
  group_by(NAME) %>%
  summarise(count = n(), mean.density = mean(density))
```

```
## Source: local data frame [3 x 3]
##
##          NAME count mean.density
##         (chr) (int)        (dbl)
## 1 Gaza Strip     2        78.92
## 2  Hong Kong     2        28.07
## 3  Singapore     2        26.25
```

You will note that we skipped naming the data object in all of functions. The piping operator means it is implicit what the data being passed on is, so there is no more need to explicitly name it.

In keeping with the piping logic, we can also use a rarely used R assignment operator: `->`. We can add it at the end of the pipe/chain and point it to the new object's name. Of course you can also start the way we have been starting all along, and assign in the standard direction `<-` if you prefer.

A whole set of piped functions can easily be wrapped up in a function:

```
FunMyPipe <- function(x) {
  x %>%
    sqrt %>%   # square root
    mean %>%   # mean
    "*" (100)  # multiplication - a bit awkward, true
}
# Test it out on a short vector
FunMyPipe(1:10)
```

```
## [1] 224.7
```

6.6 PRACTICAL - Piping Population Pyramids

```
## 4.0 Practice piping on population2010
###########################################################
```

Using the `population2010.csv` file find the answers to the following:

- How many 20 year-old males were there in Tanzania in 2010
- Which country has the lowest total population?
- In which country do women outnuber men in the most age groups?

```
## 4.1 Write a function to extract population pyramid data
################################################################################
```

Write a function that uses piping and `dplyr` functions to do the following:

- the input is the FIPS country code (if you're not sure which code means which countre, chek the list here: https://en.wikipedia.org/wiki/List_of_FIPS_country_codes)
- from `tidy.population2010` extract the data for that country
- remove variables for the area
- create a new variable grouping the ages into 5 year age groups (use `cut(x, 20)`)
- find the sum of the population for each age group and gender combination (use `group_by`!)
- don't forget to `ungroup` the data before the next step!
- create a new variable representing the proportion of the total population in each age/sex combination (* 100)
- return this table, which should have 40 rows and 4 columns.

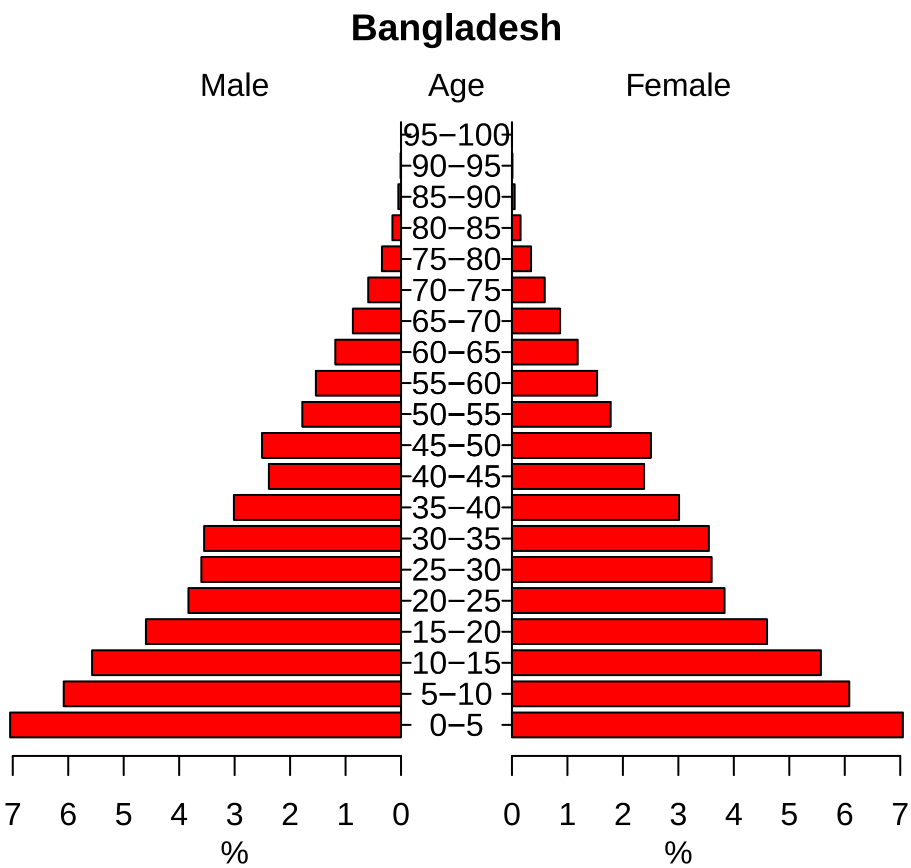
```
## 4.2 Write a function to draw a population pyramid plot
################################################################################
```

We will use the `plotrix` package for this, although you are free to experiment with drawing your own pyramid plot - you will have much more control than using the default `pyramid.plot()` function. Have a look at the help documentation for this function. In particular note that you need to provide it two vectors, `lx` and `rx` for the population sizes.

Write a function that:

- takes as it's input the output from your previous function (the 40x4 table)
- creates the `lx` and `rx` vectors
- **IMPORTANT** again you will have a data.frame as the result. use `as.matrix` on `lx` and `rx` so they can be used in the plot
- calls pyramid.plot(lx, rx) as well as any other arguments you may want to add. (in particular you may want to add `labels=`). One way of creating them is `paste(seq(0,96, 5), seq(5,100,5), sep="-")`, but you can also use the values from the age group variable.

```r
FunPlot(FunPopClean("BG"), "Bangladesh")
```




```
## [1] 5.1 4.1 4.1 2.1
```