Reproducable Research with R

Information presented is “advertising”,

Research is full software environment, code and data that produced results.

When advertising is separated from research – difficult to reproduce the results

Data gathering, statistical analysis and presentation of results

Research results are replicable if there is sufficient information for researchers to make same findings using same procedures.

the data and code used to make a finding are available and they are

sufficient for an independent researcher to recreate the finding.

Knitting the data and source code together with presentation documents.\

Research finding have to be verified through replication

Reproducibility is minimum – when obtaining a different sample is not possible – researchers should be able to reproduce the research

A study can be reproducible and still be wrong.

The goal of reproducible research is to have more impact with our research

Reproducible research involves: 1. Reproducible research environment and 2. Reproducible research publisher.

What is data science – mix of statistics and computer science.

Form hypotheses about date and examine in a structured fashion.

Datafication = data products – research gate – create a feedback loop – ex. Internet shopping

Ethical implications for data

Data gathering without permission

Use of datasets that were not meant to be used together

Once we datafy we can transform their purpose into value

Skills to create data scientist

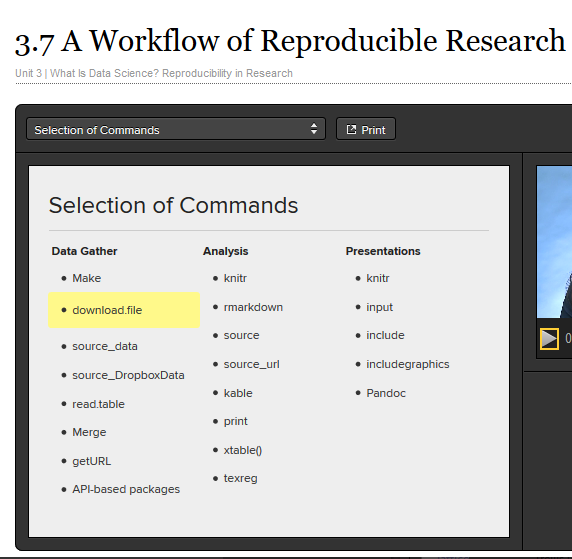
Hacking skills – finding workarounds to get necessary information

Data science task

* Collecting, cleaning, munging data
* Using persistence, statistics and software engineering skills
* Understanding biases in data
* Debugging output from code
* Combining viz and data sense
* Find patterns, build models, algorithms

3 steps in empirical research project

Data gathering, data analysis, and results presentation



Quantitative

Background information that motivates the theory

Reproducible theory also

Citations and literature search

Reproducible theory and computational results

Tips

Document everything

How data gathered, analyze,and present

R – what versions or packages were used, record session

Sessioninfo()

Use text files

Use simplist file format possible – not word or xcel, csv, asci, plain text are ‘in’

Data in text files, analysis in r and presentation in .md or .tex filed

Can be read by any software that reads text files – research can be used longer

Easy to search and manipulate – github

Human readable

Someone who has not worked on the project will be able to understand. Over-comment – codes design and purpose

Explain what code does, what graphics are meant to display

# - anything after # is comment

Write comment before block of code.

Comment any code that is ambiguous – know what you intended to do

Comment header – description, date, name, contributer, dependency on other files, version information

Tie files together

Relationship needs to be documented

Commands link files together

Table 2.1 in gandrud text has key commands for tying files together.

Have a plan

Limit content to any particular file – one file should not contain all code

Files should be modular

Reproducability needs to be part of plan from the beginning

Apply to scholarly research and private sector projects

Using R

Best to use text editor to “compose” script

Use notepad

File / new script – r has text editor too

Source or scan commands help input code from text

R session – creates workspace

Objects()

Can change working directory in R

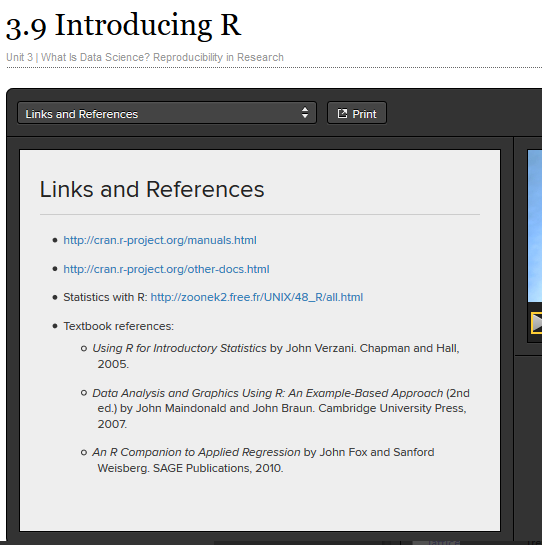
When quit R asked to save the workspace – generally not recommended.

Recommend to copy to text editor and save comments

shortcuts

Command history() – list of all commands used

Arrow keys – recall earlier scripts



Getting started in R

From desktop – right click to open R and run as administrator – need to be admin in order to download packages.

Also open notepad (text editor)

Getting help – type ?function ex; ?mean – pop up help file

If don’t know name of function - ??”average” – go to internet and find search of function / topic

Have to install and load packages

Calculator

Functions

Mean – mean

Summary – 5 number summary + mean

All data vectors have a class

Class (var) – numeric, logical (t/f), string – know class to determine if function will work

Data vectors need to be the same class

Examples

Grades

Mean(grades)

Deviations <- grades-mean(grades)

Squared deviations <- deviations^2

Sum <- sum(sq dev)

Calc <- sqrt(sum/(length(sqdev)-1))

SD() – same answer

Sessioninfo()

Save work – copy into notepad ---- text file with r code – clean it up – remove errors and comment data

Indexing

Grades

Grades [1]

Grades [2:7]

Grades [c(2,4,8)]

Grades [-2] – all except what is specified

To replace

Gpa.section.a[5] <- 3.80

To add

C(gpa.section.a, 3.6 2.8)

Statistical data set = data frame

Type in dataset name to load data

Ex stackloss

Not great idea – don’t know how big dataset is

Dim(stackloss) – returns rows x columns

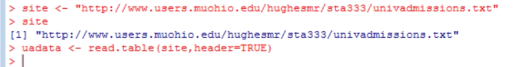
Str(stackloss) – tells class, dimensions, list variables and class

1st step to get familiar with data set

Head(stackloss) – gives 1st 6 lines or head(stackloss,10) will return 10

Tail (…) – last 6 lines, etc.

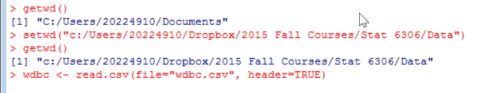
\



Names (file) – gives variable names

Getwd() – get working directory

Setwd(“c:/lllll/lllll/lllll)



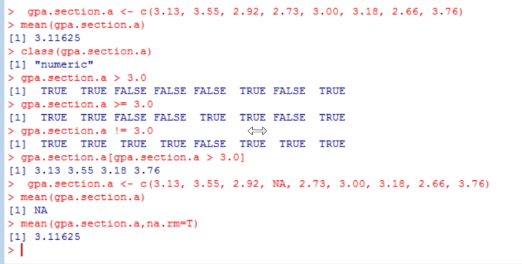
Logical expression

Grade >3.6 = returns true false

< <= >>= != --- not equal to

Can use logical values to subset

What about missing values – R will insert “NA” – will get NA for quantitative calculations





Which(is.na(gpa) – return which values

NaN – undefined values – not mathematically permissible

1/0 – Inf – we have divided by 0 or close to zero – getting

Chapter 2

3 stages of research

Data gathering

Data analysis

Presentation

Figure 2.1 – graphic to represent commands used for the above

Tips for reproducible research

1. Document everything

Record session info

Sessioninfo()

1. Everything in a text file

Should use simplest file format available (txt)

Helps future proof your research

1. All file should be human readable

Treat files as if someone who has not worked on the project will want to understand some day

Comments begin with #

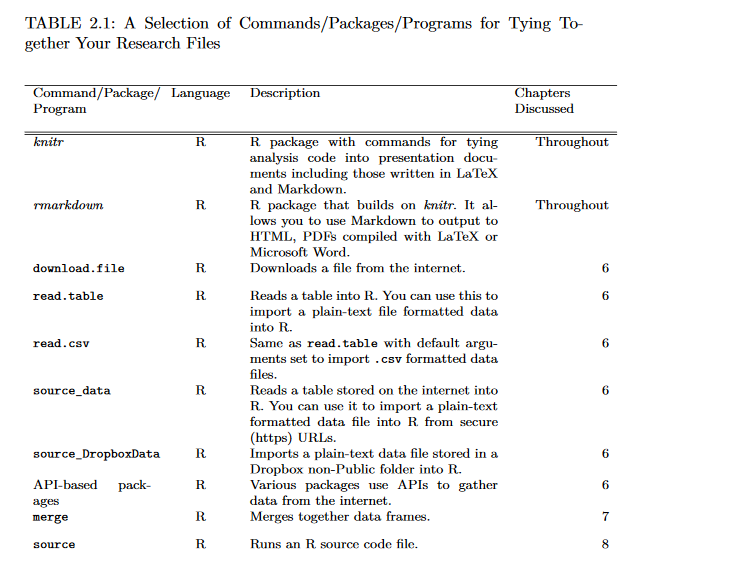
Guidelines – write comment before block of code to describe what it does, and also any ambiguous code

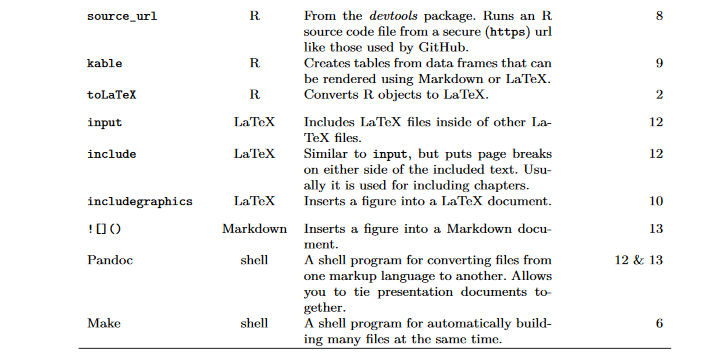
Use style guides

1. Explicitly tie files together

Use tie commands

1. Plan to store, organize and make files available





Ch 3

R is object – oriented language

They are things like a vector of numbers, data set, a word, a table

R is focused on doing actions on objects

NA – means “missing”

Class() – returns “character”, etc.

Vectors make up dataframes

Cbind(one,two) – combines vectors together

Dataframe(one,two) – allows to combine with different classes ex: numeric and character

Names(dataframe) – displays vector names

$ - component selector

Asynch

Args(sd) – args function will return arguments associated with a command

Use if you’re not sure what values can go into it

Creating R functions

Function (argument-list) {

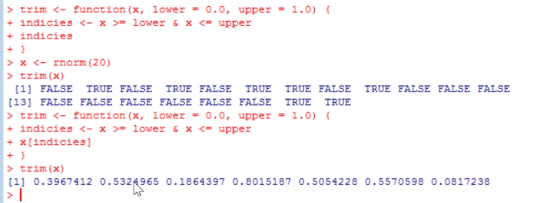
expressions

value

}









Looping helps computer do same thing over and over again

“for loops”

Getwd()

Shows directory saving to

Setwd(“c:\....”) find full directory path in documents file

Change direction of the slash in windows or add two slashes

List.files() – will list files in directory

Dir.create(“exampledir”) – will create in working directory

File.create(“exampledir/code.r”)

Relative to the current working directory

Path = paste(getwd(), “[\\Exampledir](file:///\\Exampledir)”, sep=””)

Setwd(path)

Getwd()

List.files()

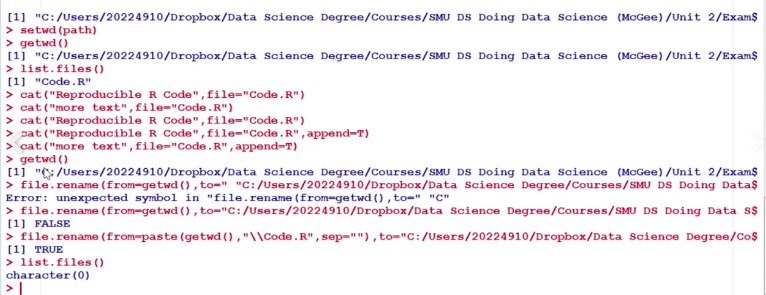
Cat(“reproducible r code”, file = “code.R”)

Cannot type cat again – will overwrite existing data

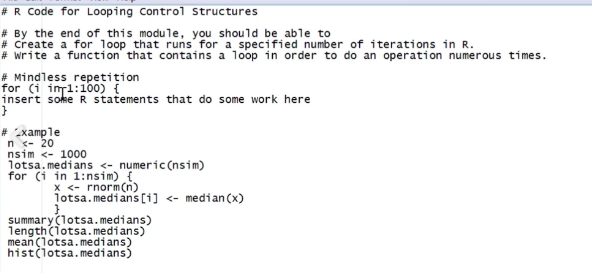
Cat(“more text”, file = “code.R”, append = T)

File.rename

Can be used to move a file



Looping control structures

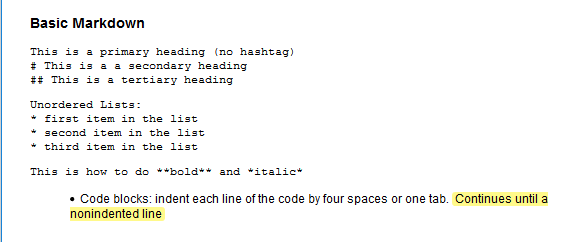


Markdown – mostly human readable

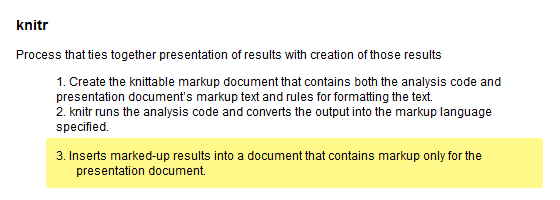
Button in r studio to get basic help –

“introduction to markdown”

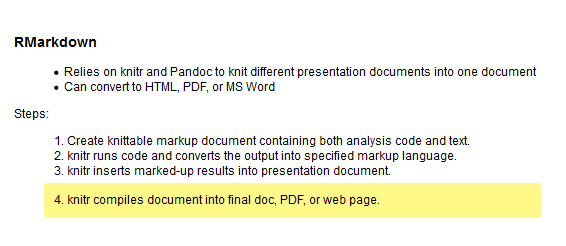
R – markdown has place for code.

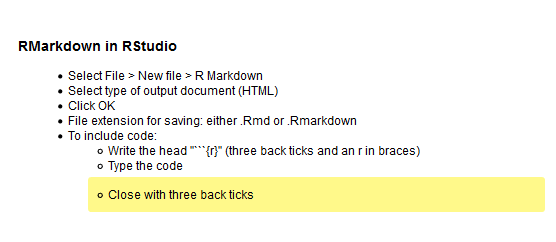


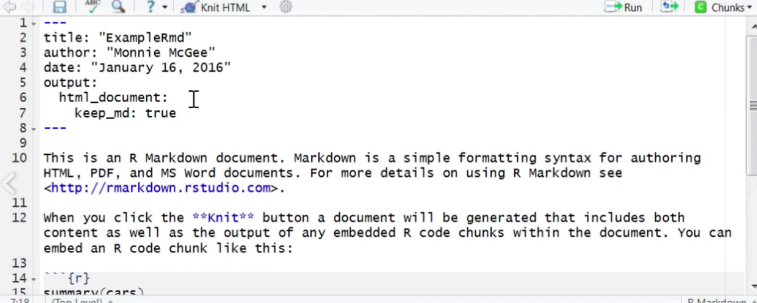
Knitr is backend to markdown



R markdown – knitr and pandoc







Keep markdown for posting to github

