Week 1 Demonstration

Data Preprocessing: From Raw Data to Ready to Analyse

Dr. Anil Dolgun

28/02/2018

1 / 20

Image credits: RMIT University, https://flic.kr/p/vArsRW

2 / 20

Course coordinator

- Dr. Anil Dolgun
- Lecturer in School of Science, Mathematical Sciences, RMIT University
- E-mail: anil.dolgun@rmit.edu.au
- Office: Building 8, Level 9, Room 23 (Contact for appointment)
- LinkedIn: https://au.linkedin.com/in/anildolgun

Get Started

Course Orientation

- This course assumes you have a working knowledge of basic mathematics and familiarity with computers.
- - Please read: You must be logged into your RMIT Student Google account to access this document.
- : http://rare-phoenix-161610.appspot.com
- : Wednesdays 18:30 - 21:30 in 080.02.02
  - Announcements and Questions (~ 5-10 mins)
  - Demonstration (~1 hr)
  - Class Activities (~ 1 hr, exercises on Class Worksheets)
  - Supervised self-directed learning (~ 1 hr, work on module skill builders and/or assignments)
- :
  - Skill Builders
  - DataCamp modules
- (see Course Information Pack)
- : MATH2349 has been designed to run online. Classes are recorded (Canvas - EchoCenter), attendance is not compulsory.
- Course is under development so materials will become available as the semester progresses.

Slack - Course Communication

- I am trialing a new course communication tool this semester - Slack

https://math2349.slack.com

- Sign-up here https://join.slack.com/math2349/signup
- This will be the go-to-place for all course communication.
- Please only email me if you have personal academic matters to discuss.

DataCamp Online Courses

MATH2349 Data Preprocessing course is supported by DataCamp for the Classroom initiative.

- During this semester, you will have free access to DataCamp learning modules.
- I have selected specific modules that you will need to complete as a skill builder.
- and follow the instructions to sign-up MATH2349 Data Preprocessing group on DataCamp.

- You will have 6 months of FREE access to the full DataCamp course curriculum (>250 hours).
- Access to premium courses (i.e., R, Python and SQL courses).
- You can participate in leaderboards and private discussion forums with your fellow classmate.
- You may also complete other online courses that you are interested as they will help you with your other studies.

Course Feedback

- Official CES feedback comes too late.
- To further develop new ideas and/or improve the current curriculum I need feedback. I take feedback seriously.
- Please fill out this form (Available from Tools --> Course Feedback on the website) at any stage of the semester.

- Often, I can solve issues right away :)
- Be nice and realistic!

Assessment

- Course assessment is comprised of the following:
  - (60%): Three assignments spread throughout the course. All assignments are due on Sundays 11:59pm AEST, unless otherwise stated.
  - (40%): Supervised, paper-based exam during the exam period. Details to follow later in the semester.

- Assignment 1 details are available.

Module 1 Basics : What is Data Preprocessing?

What is Data Preprocessing?

- is a process and the collection of operations needed to prepare all forms of untidy data (incomplete, noisy and inconsistent data) for statistical analysis.

We will define 5 major tasks for data preprocessing framework, namely : , , , and .

What will you learn in this course?

- Module 1: Basics: What is DP?
- Module 2: Get
- Module 3: Understand
- Module 4: Tidy & Manipulate
- Module 5: Scan: Missing Values
- Module 6: Scan: Outliers
- Module 7: Transform
- Module 8: Special Operations

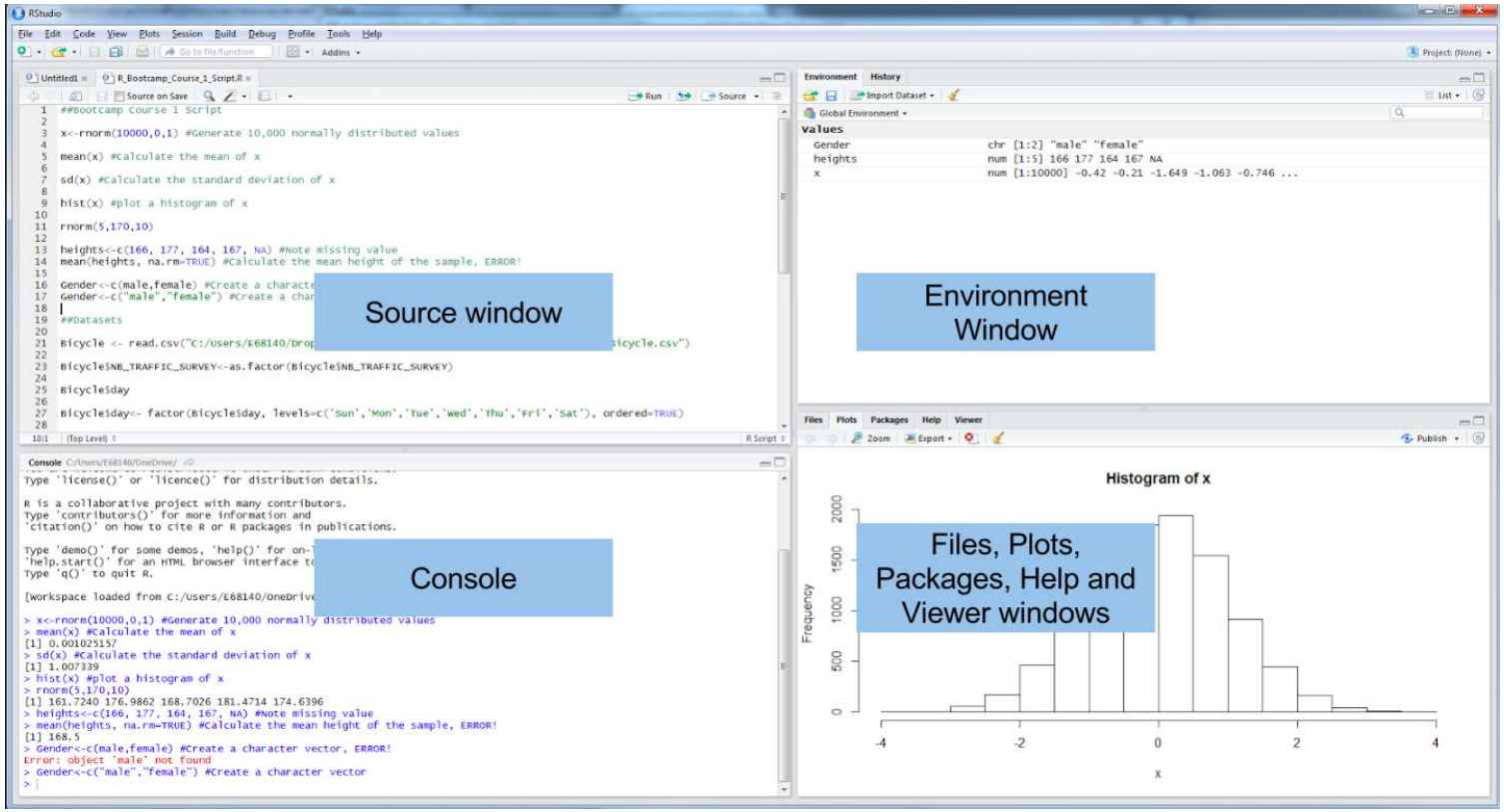
- : Open Source R/RStudio*
-
  - Class worksheets
  - Data challenges
  - Skill Builders
  - Assignments

*: Including Base R functions, , , , , , packages and many others.

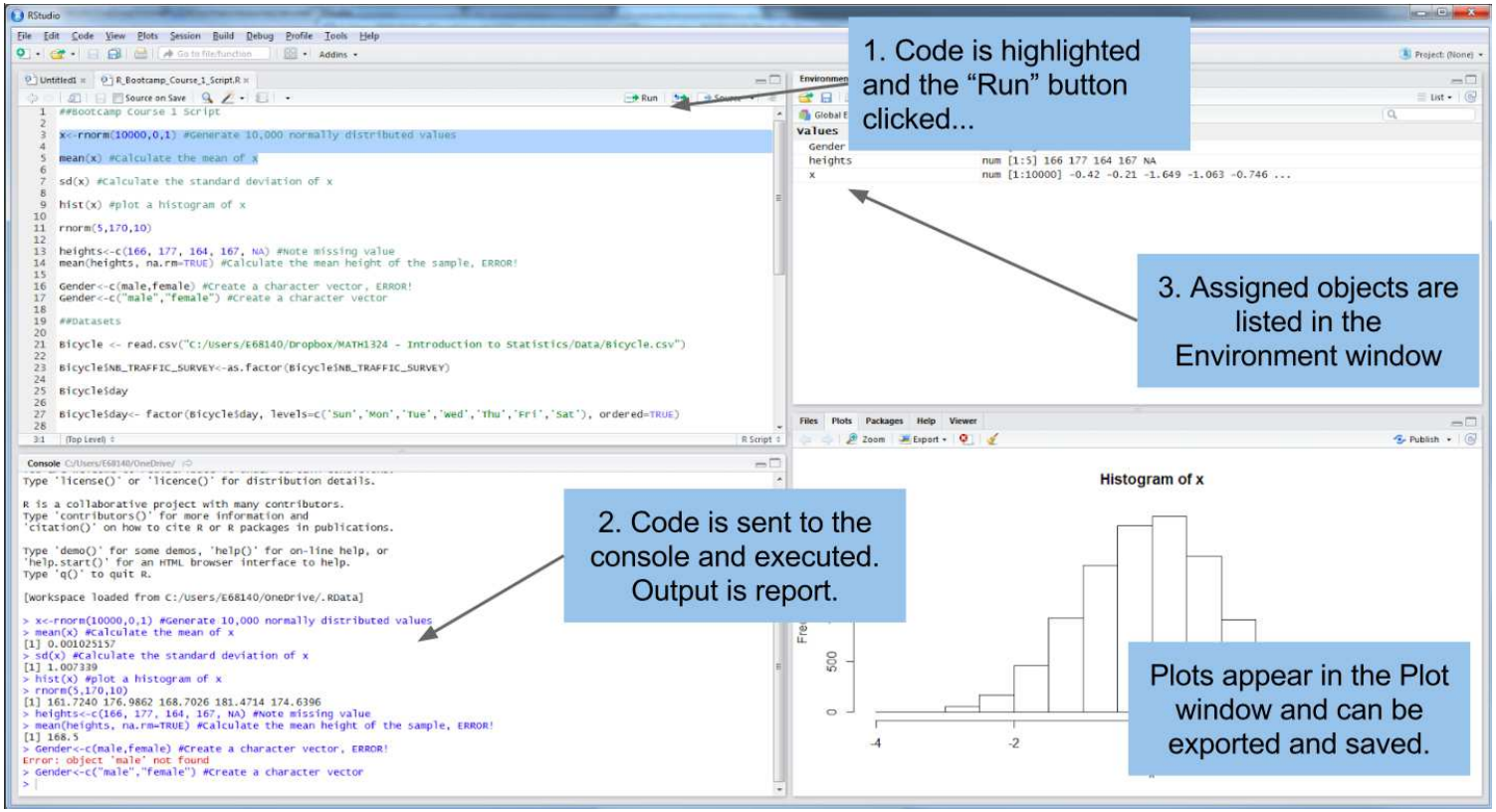
R and RStudio Quick Overview

- R is a free programming language and environment for statistical computing - https://www.r-project.org/
- Why learn R?
    - Recognised across industries
    - Promotes coding and computational skills
    - Provide access to the world's largest and most comprehensive library of statistical functions (not available in other commercial statistical softwares)
    - Powerful and grows with you
    - Works on all major operating systems
    - R and RStudio can be used in combination to create new functions and statistical programs, build dynamic and interactive reports, dashboards, websites, slideshows and statistical web applications and all for….
    - free!
- RStudio is a free integrated development environment for R and makes using R a lot easier and more efficient - https://www.rstudio.com/.
- RStudio requires R to be installed.

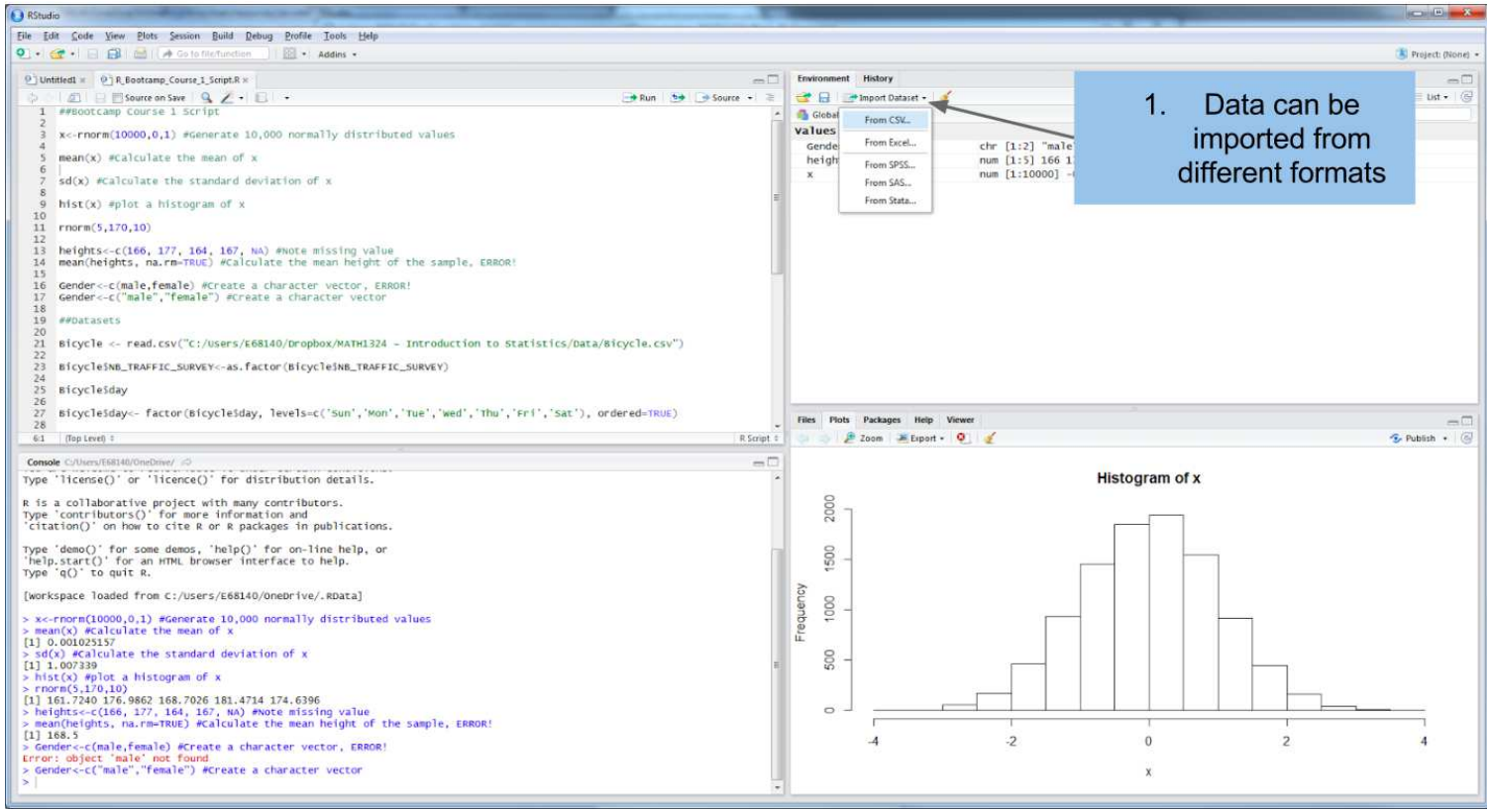
R and RStudio Quick Overview Cont.1



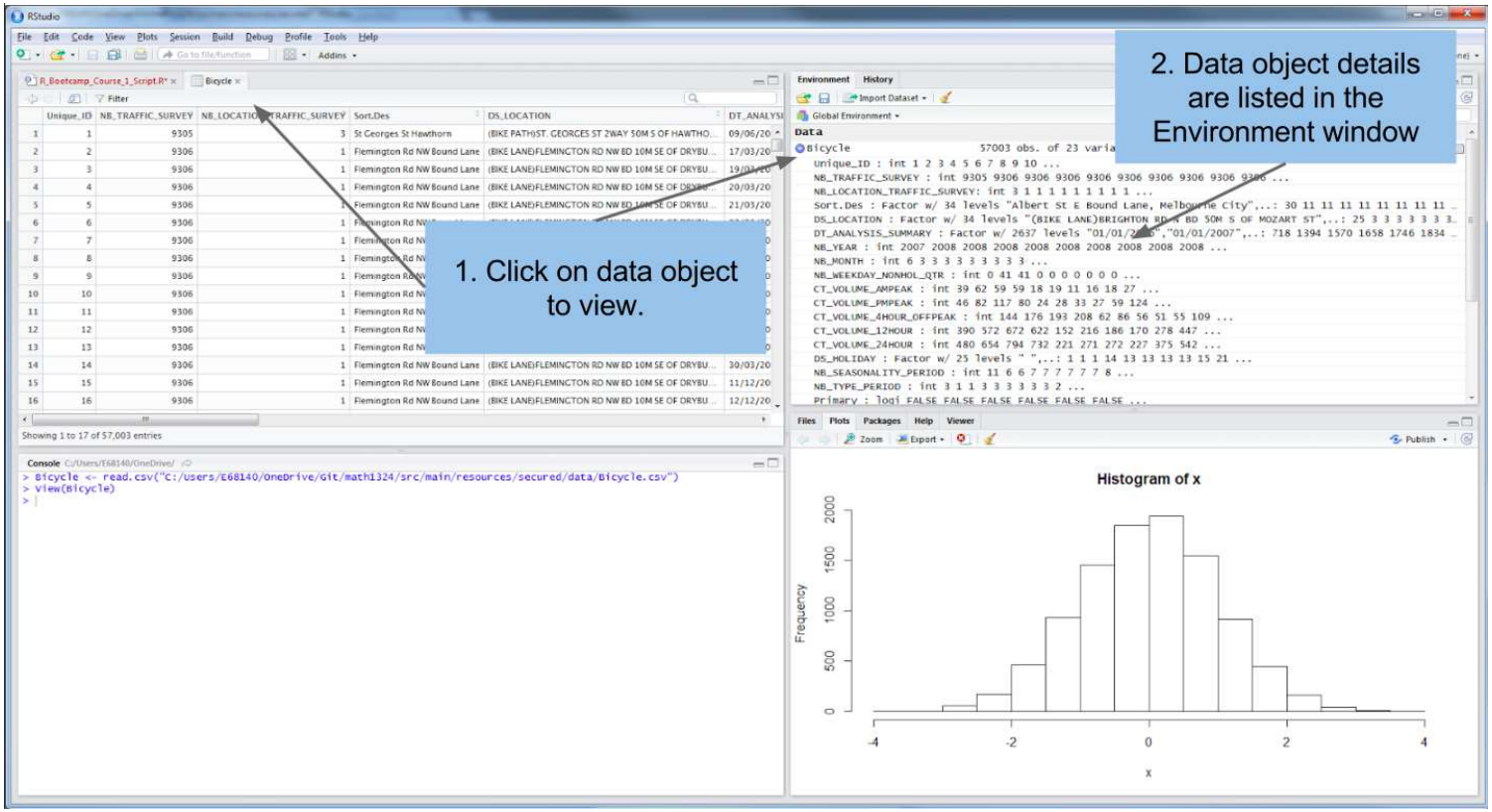
R and RStudio Quick Overview Cont.2



R and RStudio Quick Overview Cont.3



R and RStudio Quick Overview Cont. 4



Installing and Loading Packages

- Packages are collections of related functions. Comprehensive R Archive Network (CRAN) lists over 10,000 available packages!
- Packages are the reason why R is so powerful.
- Packages need to be installed and loaded into an R session.

- It is a good idea to include the option as many packages require other packages to run. This option checks and installs dependent packages where required.
- Once a package is installed, it can be loaded into an R session in order to make its functions available.

- You will .
  Always start your scripts, notebooks or markdown files by loading all the packages you will need.

What do you need to know by Week 1

- Read through the Course Information Pack
- How to access the course Canvas shell through myRMIT
- How to access our course website
- How to access Slack
- How to access MATH2349 on DataCamp
- Learn how to install R and RStudio, know the overview of the RStudio interface refer to Dr. James Baglin's R Bootcamp.
- Know how to install and load R packages (See Module 1 notes)
- Know how to get further help for R statistical programming language (refer to Module 1 notes)
- Don’t panic. R has a slow learning curve, but you will get heaps of practice in this course!

Class Worksheet

- Working in small groups, complete the following class worksheet

Week 1 Class Worksheet

- Once completed, feel free to work on your Assignment and/or Skill Builders

Return to Data Preprocessing Website