

# 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>

# Course coordinator

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



# Get Started

4 / 20

# 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.
- [follow the instructions to sign-up MATH2349 Data Preprocessing group on DataCamp.](#) and
- 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.

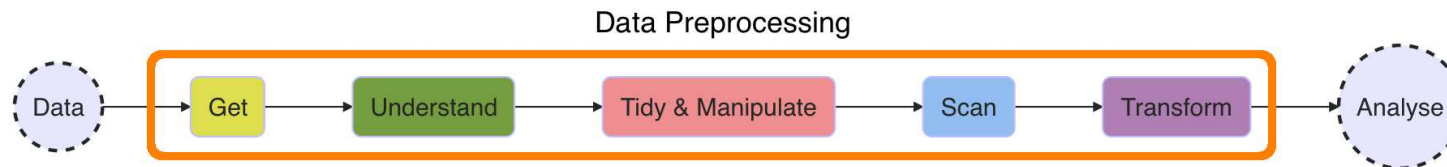
# Module 1 Basics : What is Data Preprocessing?

10 / 20

# 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 .





# 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

The screenshot displays the RStudio environment with four main windows highlighted by blue boxes:

- Source window:** Contains R code for generating random values, calculating mean and standard deviation, plotting a histogram, and reading a CSV file.
- Environment window:** Shows the current environment with variables like 'Gender', 'heights', and 'x'.
- Console:** Displays the output of the R code, including the execution of 'rnorm', 'mean', 'sd', 'hist', and 'read.csv'.
- Files, Plots, Packages, Help and Viewer windows:** Shows a histogram of the variable 'x'.

The R code in the Source window is as follows:

```
1 ##bootcamp course 1 script
2
3 x<-rnorm(10000,0,1) #Generate 10,000 normally distributed values
4 mean(x) #calculate the mean of x
5 sd(x) #calculate the standard deviation of x
6
7 hist(x) #plot a histogram of x
8
9 rnorm(5,170,10)
10
11 heights<-c(166, 177, 164, 167, NA) #note missing value
12 mean(heights, na.rm=TRUE) #Calculate the mean height of the sample, ERROR!
13
14 Gender<-c(male,female) #create a character vector, ERROR!
15 gender<-c("male","female") #create a character vector
16
17 ##Datasets
18
19 bicycle <- read.csv("C:/Users/E68140/OneDrive/My Documents/Bicycle.csv")
20
21 bicycle$NB_TRAFFIC_SURVEY<-as.factor(bicycle$NB_TRAFFIC_SURVEY)
22
23 bicycle$day
24
25 bicycle$day<- factor(bicycle$day, levels=c('Sun','Mon','Tue','Wed','Thu','Fri','Sat'), ordered=TRUE)
26
27
28
```

The Environment window shows the following variables:

Variable	Class	Length	Values
Gender	chr	1:2	"male" "female"
heights	num	1:5	166 177 164 167 NA
x	num	1:10000	-0.42 -0.21 -1.649 -1.063 -0.746 ...

The Console shows the following output:

```
> x<-rnorm(10000,0,1) #Generate 10,000 normally distributed values
> mean(x) #calculate the mean of x
[1] 0.001025137
> sd(x) #calculate the standard deviation of x
[1] 1.007339
> hist(x) #plot a histogram of x
> rnorm(5,170,10)
[1] 161.7240 176.9862 168.7026 181.4714 174.6396
> heights<-c(166, 177, 164, 167, NA) #note missing value
> mean(heights, na.rm=TRUE) #Calculate the mean height of the sample, ERROR!
[1] 168.5
> Gender<-c(male,female) #Create a character vector, ERROR!
Error: object 'male' not found
> gender<-c("male","female") #create a character vector
>
```

The Files, Plots, Packages, Help and Viewer windows show a histogram of the variable 'x' with the title 'Histogram of x'. The x-axis is labeled 'x' and ranges from -4 to 4. The y-axis is labeled 'Frequency' and ranges from 0 to 2000.

# R and RStudio Quick Overview Cont.2

The screenshot shows the RStudio environment with four main panels: Source, Environment, Console, and Plots. Annotations highlight the following features:

- 1. Code is highlighted and the "Run" button clicked...**: Points to the 'Run' button in the Source panel toolbar.
- 2. Code is sent to the console and executed. Output is report.**: Points to the Console panel showing the execution of R code.
- 3. Assigned objects are listed in the Environment window**: Points to the Environment panel showing the objects created in the workspace.
- Plots appear in the Plot window and can be exported and saved.**: Points to the Plots panel showing a histogram of the variable 'x'.

**Source Panel Code:**

```
##bootcamp course 1 script
1 x<-rnorm(10000,0,1) #Generate 10,000 normally distributed values
2
3
4 mean(x) #calculate the mean of x
5
6 sd(x) #calculate the standard deviation of x
7
8 hist(x) #plot a histogram of x
9
10
11 rnorm(5,170,10)
12
13 heights<-c(166, 177, 164, 167, NA) #note missing value
14 mean(heights, na.rm=TRUE) #calculate the mean height of the sample, ERROR!
15
16 Gender<-c(male,female) #Create a character vector, ERROR!
17 Gender<-c("male","female") #Create a character vector
18
19 ##Datasets
20
21 Bicycle <- read.csv("C:/Users/E68140/Dropbox/MATH1324 - Introduction to Statistics/data/Bicycle.csv")
22
23 Bicycle$INB_TRAFFIC_SURVEY<-as.factor(Bicycle$INB_TRAFFIC_SURVEY)
24
25 Bicycle$day
26
27 Bicycle$day<- factor(Bicycle$day, levels=c('Sun','Mon','Tue','Wed','Thu','Fri','Sat'), ordered=TRUE)
28
```

**Environment Panel:**

Global Environment	values
Gender	num [1:5] 166 177 164 167 NA
heights	num [1:10000] -0.42 -0.21 -1.649 -1.063 -0.746 ...
x	

**Console Panel Output:**

```
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[workspace loaded from c:/users/E68140/OneDrive/.RData]

> x<-rnorm(10000,0,1) #Generate 10,000 normally distributed values
> mean(x) #calculate the mean of x
[1] 0.001025157
> sd(x) #calculate the standard deviation of x
[1] 1.007339
> hist(x) #plot a histogram of x
> rnorm(5,170,10)
[1] 161.7240 176.9862 168.7026 181.4714 174.6396
> heights<-c(166, 177, 164, 167, NA) #note missing value
> mean(heights, na.rm=TRUE) #calculate the mean height of the sample, ERROR!
[1] 168.5
> Gender<-c(male,female) #Create a character vector, ERROR!
Error: object 'male' not found
> Gender<-c("male","female") #Create a character vector
>
```

**Plots Panel:** Histogram of x

# R and RStudio Quick Overview Cont.3

The screenshot displays the RStudio environment with the following components:

- Script Editor:** Contains R code for generating random values, calculating mean and standard deviation, plotting a histogram, and reading a CSV file. The code includes comments and error messages.
- Console:** Shows the execution output of the script, including the workspace loaded from a local path and the results of the calculations. It also displays error messages for the 'Gender' variable creation.
- Environment:** Lists the objects in the global environment, including 'chr', 'num', and 'x'. A dropdown menu for 'Import Dataset' is open, showing options like 'From CSV...', 'From Excel...', 'From SPSS...', 'From SAS...', and 'From Stata...'. A blue callout box points to this menu with the text: "1. Data can be imported from different formats".
- Plots:** A histogram titled 'Histogram of x' is displayed, showing the frequency distribution of the variable 'x'. The x-axis ranges from -4 to 4, and the y-axis (Frequency) ranges from 0 to 2000.



# R and RStudio Quick Overview Cont. 4

The screenshot displays the RStudio environment with the 'Bicycle' data object loaded. The Environment window on the right shows the details of the 'Bicycle' object, including its structure and variable types. A blue box highlights the text: "2. Data object details are listed in the Environment window".

The main editor window shows a table of data with columns: Unique\_ID, NB\_TRAFFIC\_SURVEY, NB\_LOCATION, SortDes, DS\_LOCATION, and DT\_ANALYSIS. A blue box highlights the text: "1. Click on data object to view." with an arrow pointing to the 'Bicycle' object in the Environment window.

The Console window at the bottom shows the following commands:

```
> B1cycle <- read.csv("C:/Users/E68140/OneDrive/Git/math1324/src/main/resources/secured/data/Bicycle.csv")
> View(B1cycle)
> |
```

A histogram titled "Histogram of x" is displayed in the bottom right corner, showing the frequency distribution of the variable 'x'. The x-axis ranges from -4 to 4, and the y-axis (Frequency) ranges from 0 to 2000. The distribution is roughly bell-shaped, centered around 0.

# 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 `install.packages(dependencies = TRUE)` 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](#)