



centre for
mathematical
modelling of
infectious diseases



Saw Swee Hock
School of Public Health

TM-CM02 Biostatistics for Public Health

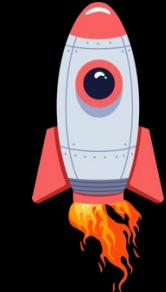
Lecture 0

Introduction to R

A practical session on data handling, cleaning and analysis for beginners

Kiesha Prem

Saw Swee Hock School of Public Health, National University of Singapore



How to survive/enjoy this class?

Statistics and programming can be fun!

Materials will be uploaded to the GitHub repo.

If you are lost, ask questions, stop me/look for Dorothy!

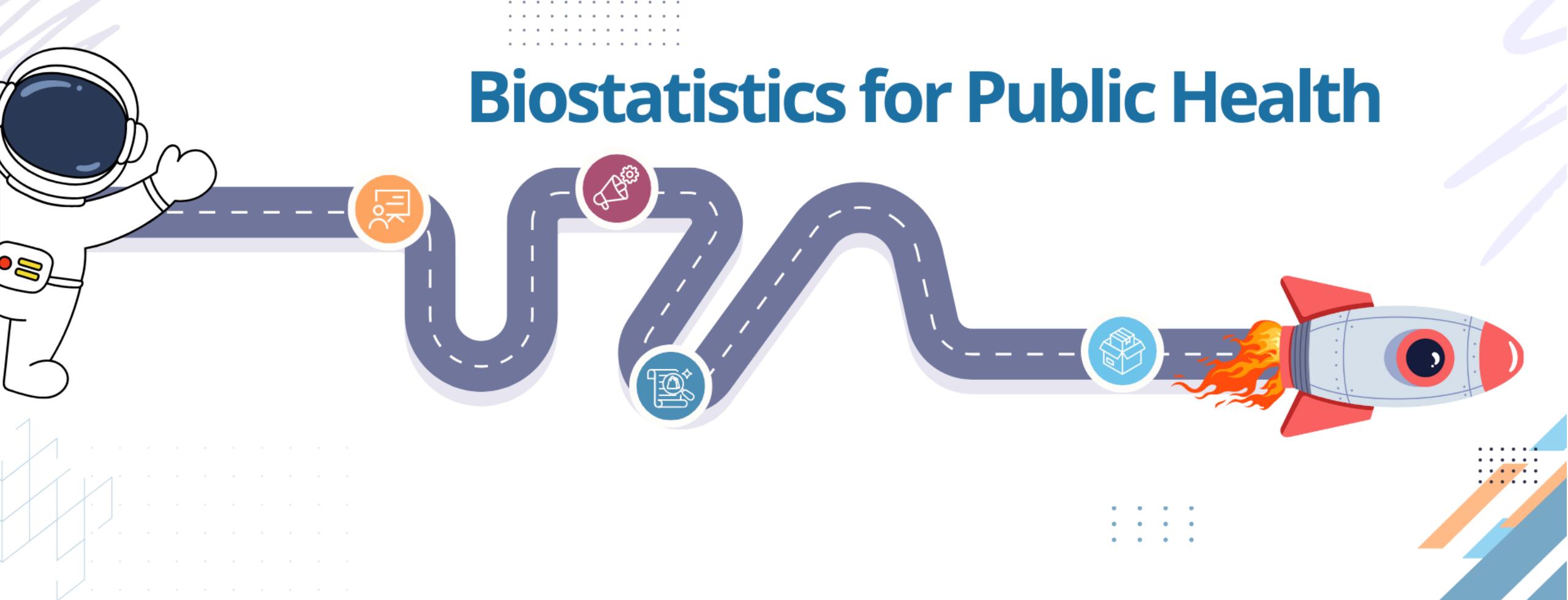
She is a **statistician**
and very nice but
please don't bully her!

Warning

If you don't ask me questions, I will ask you questions

Ask your seniors! ☺

Biostatistics for Public Health



**Share a little about
yourself**

Biostatistics for Public Health

• quizzes
• 2 assignments

Categorical data analysis

Data prep, descriptive & inferential statistics



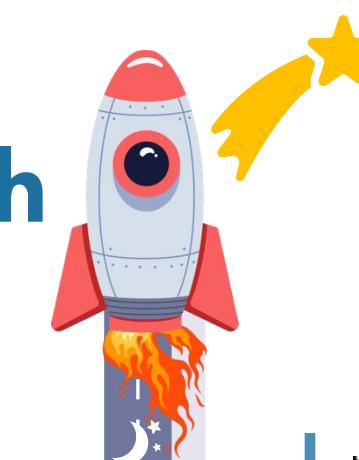
Linear regression

Data prep, descriptive & inferential statistics, linear regression



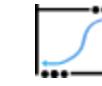
Hypothesis testing

Inferential statistics

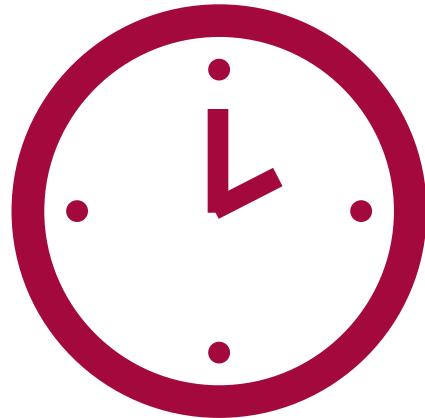


Logistic regression

Data prep, descriptive & inferential statistics, logistics regression and others if we have time

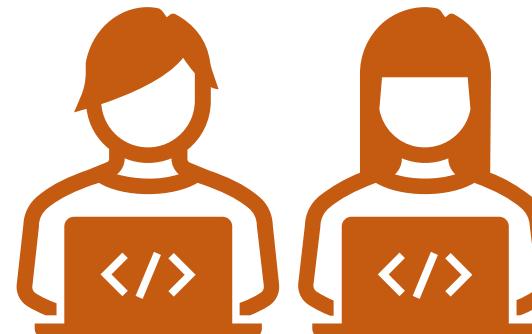


Today's class

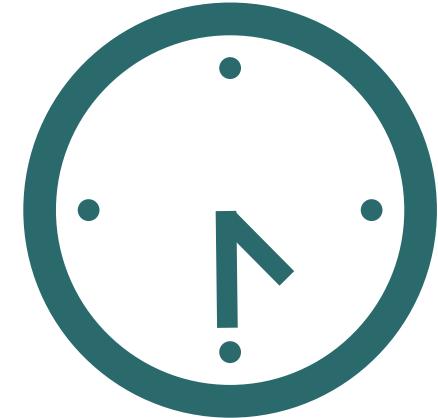


The class will start with lectures at 2 pm (There is a quiz next week!)

Today's class is special because you will need to install 3 apps and create 1 account



15-min Snack/bio break



Practicals will end at around ~430 pm



First... let's get the basics sorted

1. Decide on an email that you will link to your GitHub account (more information later)
 - a. **Choose wisely** – make sure you can receive this email and remember the password
2. Decide on a username and password for your GitHub account



kieshaprem

Hello world
this is me! ☺

3. You may need to set up a 2FA account for security purposes, e.g. on your phone, ...

Use the next 5 mins to do these activities we will use them very soon



Install R and RStudio

Nowadays, most people use the GUI RStudio to work with R.

To install RStudio, you first need to install R and then install RStudio.

Select the appropriate version of R for your operating system (Windows, macOS, or Linux).



1. Download & install R: <https://cran.r-project.org/>
2. Download & install RStudio: <https://posit.co/download/rstudio-desktop/>

This may take you about 10–20 mins with good internet

Reminder: Please install R before RStudio!

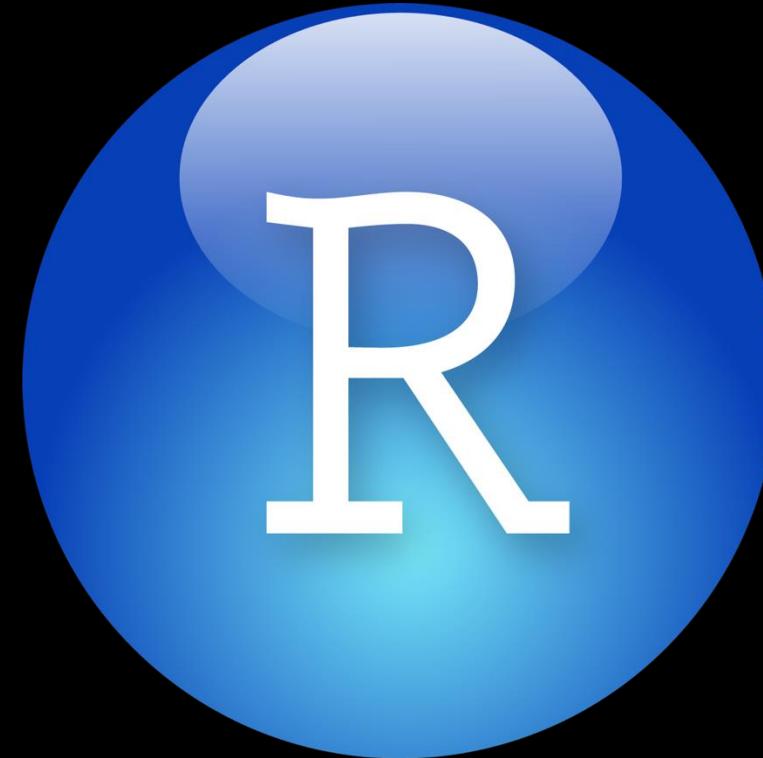


Install R





Install RStudio





Basics in R programming

Introduction to R

A practical session on data handling, cleaning and analysis for beginners

Optional for those who are familiar with R

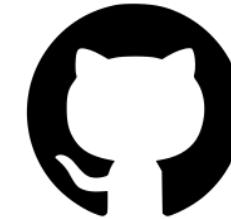
Help in R

- Use the `?function_name` or `help(function_name)` commands in the console to get help on specific R functions
- Communities like [Stack Overflow](#) are great places to find answers and ask questions if a similar question has not been asked.





Why GitHub?



One of the easiest ways to use GitHub is through GitHub Desktop without directly using Git commands:

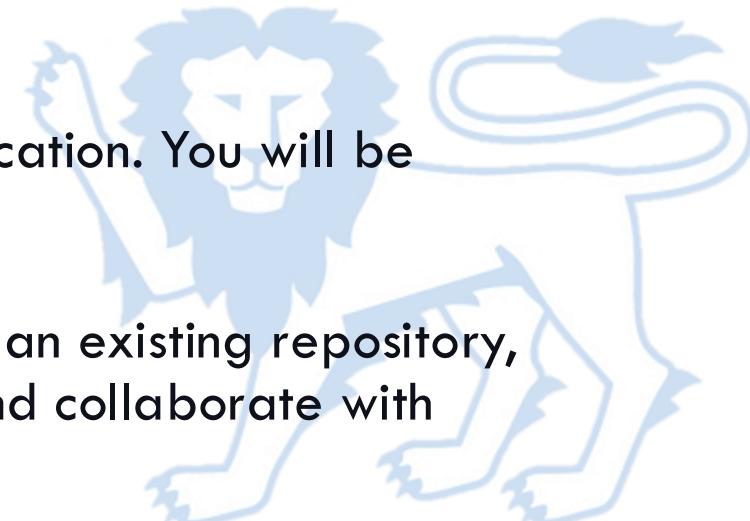
- 1. Sign up for a GitHub account:** <https://github.com/>

- 2. Download GitHub Desktop:** Visit <https://desktop.github.com/download/> and download the appropriate version for your operating system (Windows or macOS).

- 3. Install the application:** Follow the installation instructions specific to your operating system.

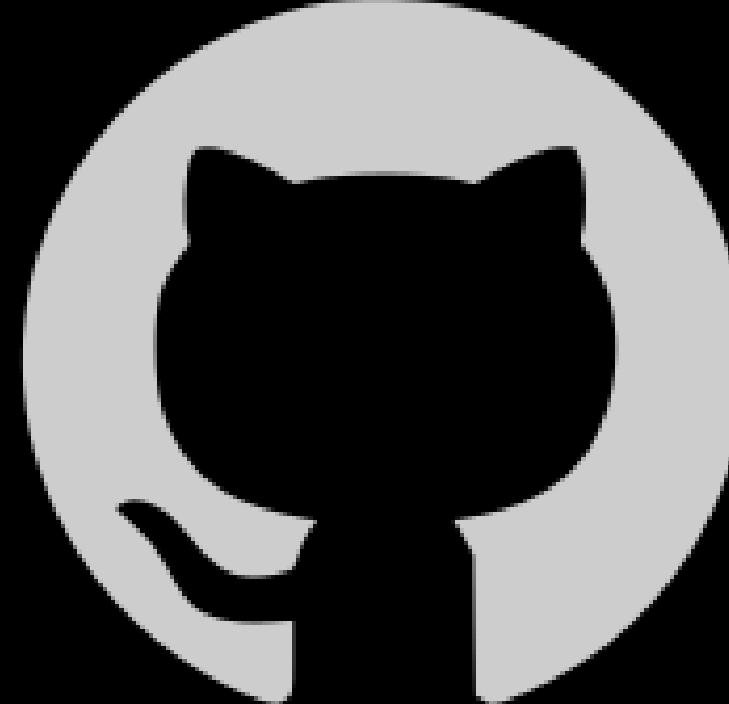
- 4. Sign in to GitHub:** Once installed, open the GitHub Desktop application. You will be prompted to sign in to your GitHub account (step 1).

5. In the next in-person practical session, we will review how to clone an existing repository, create a new repository, make changes to files, commit changes and collaborate with others.



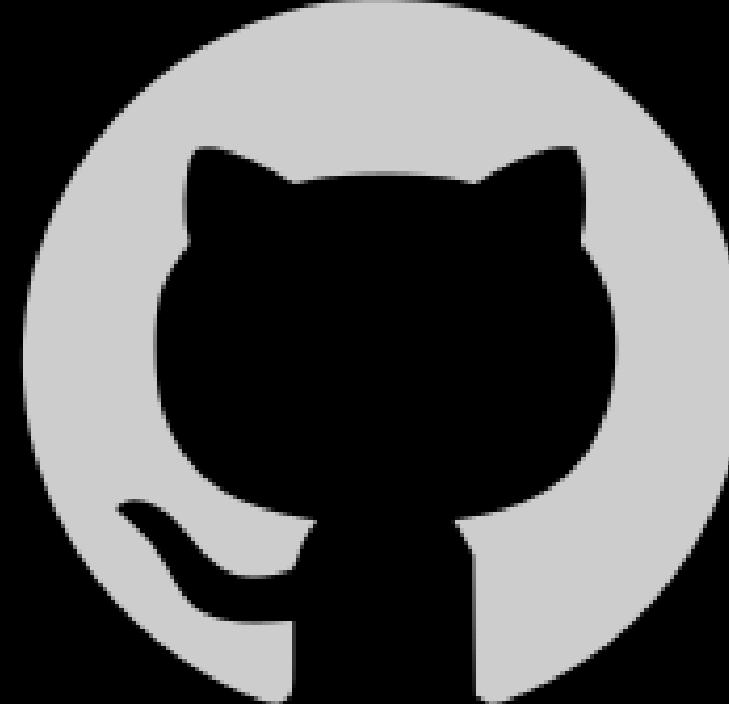


Create a GitHub account





Install GitHub Desktop





Find the class page



[tm-cm02-2025](#)

Public

TM-CM02 repository

GNU General Public License v3.0

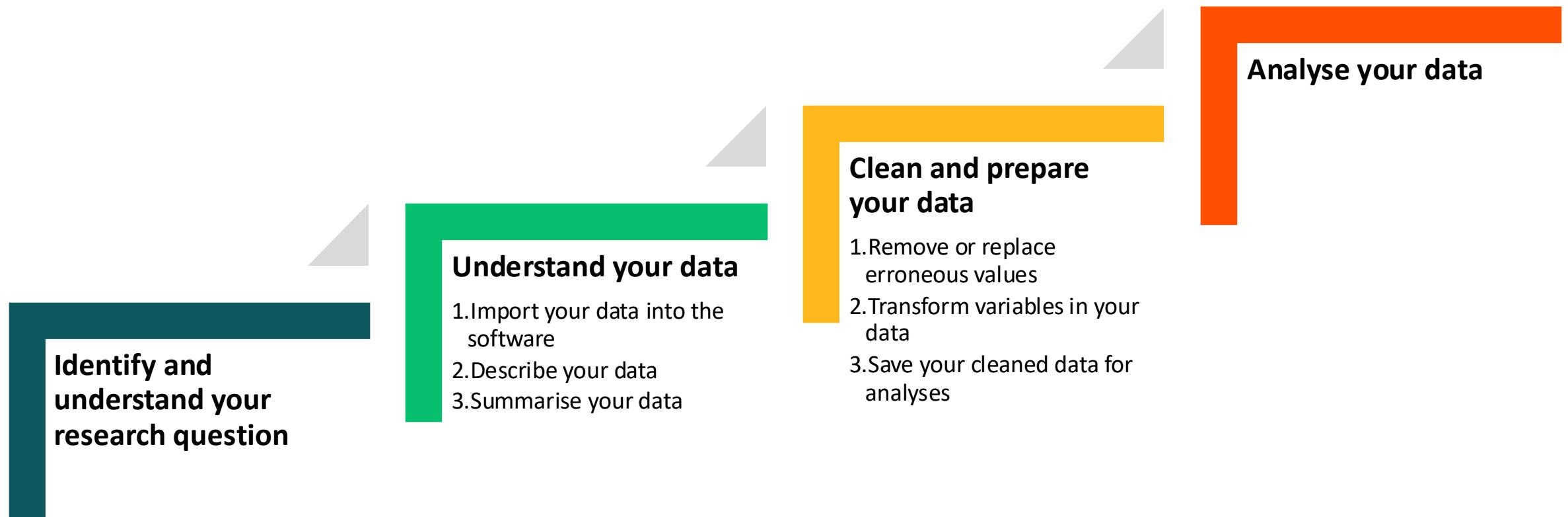
Updated 4 hours ago

15-min Snack/bio break



<https://youtu.be/iGjv-igK4ll>

Data analysis



Data analysis

Identify and understand your research question

Understand your data

1. Import your data into the software
2. Describe your data
3. Summarise your data

Clean and prepare your data

1. Remove or replace erroneous values
2. Transform variables in your data
3. Save your cleaned data for analyses

Analyse your data

Will be covered in class over the course

Overview of the study

Study Design

A community-based cross-sectional survey of 10,022 respondents aged 40 years and above in Singapore was conducted to examine the factors associated with cardiovascular risk

Disproportionate stratified sampling of ethnic groups was undertaken (only individuals of Chinese, Malay or Indian ethnicity were included)

Data collected

1. Age - in completed years (**in integers**)
2. Gender (**Male/Female**)
3. BMI – measured height and weight (**calculated to one decimal place**)
4. Ethnicity (**Chinese/Malay/Indian**)
5. Smoking status – self-reported by participants (**Daily smoker/Occasional smoker/Ex-smoker/Never smoker**)
6. LDL cholesterol – measured from fasting blood samples obtained from participants (**available up to two decimal places**)
7. Presence of cardiovascular disease– self-reported by participants as having being diagnosed by a physician (**Yes/No**)

Research Questions

1. Is gender associated with levels of LDL cholesterol?
1-a. Is this association confounded by smoking status?

2. Is BMI associated with CVD?
2-a. Is this association confounded by smoking status?

What does the data look like?

First column contains unique participant ID

Current data format : Excel

First row contains the variable names

1	A	B	C	D	E	F	G	H
	id	age	gender	bmi	ethnicity	smoke	cvd	ldl
2	1	72	Female	23.9	Indians	Never-Smoker	0	3.49
3	2	73	Female	26.2	Chinese	Never-Smoker	0	3.55
4	3	67	Female	19.9	Malays	Never-Smoker	0	3.15
5	4	65	Female	27.8	Indians	Never-Smoker	0	2.97
6	5	72	Male	22.0	Indians	Daily smoker	0	3.90
7	6	55	Female	20.9	Indians	Never-Smoker	0	2.29
8	7	72	Female	21.8	Malays	Daily smoker	1	3.92
9	8	66	Female	28.3	Malays	Never-Smoker	0	3.06
10	9	66	Male	27.5	Malays	Never-Smoker	0	3.06
11	10	62	Female	21.9	Chinese	Occasional smoker	0	3.14
12	11	67	Male	20.9	Malays	Never-Smoker	0	3.14
13	12	81	Female	11.6	Indians	Occasional smoker	0	4.51
14	13	71	Female	34.2	Malays	Never-Smoker	0	3.39
15	14	72	Male	22.5	Indians	Never-Smoker	0	3.46
16	15	63	Male	23.8	Malays	Never-Smoker	0	2.82

10001	10000	68	Female	25.6	Malays	Never-Smoker	0	3.21
10002	10001	4	Female	24.9	Malays	Never-Smoker	0	3.09
10003	10002	141	Male	22.8	Indians	Never-Smoker	0	2.79
10004	10003	150	Female	19.6	Indians	Never-Smoker	0	2.76
10005	10004	2	Female	21.5	Chinese	Never-Smoker	0	3.55
10006	10005	82	Female	22.8	Malays	zsmoker	1	2.30

Total number of participants is 10,022.

10018	10017		Male	37.7	Indians	Never-Smoker	1	
10019	10018	55		22.4	Indians	Never-Smoker	1	
10020	10019	60	Female	31.9		Occasional smoker	0	3.00
10021	10020	66	Female	29.7		Never-Smoker	0	3.03
10022	10021		male	31.2	Malays	Occasional smoker	0	3.66
10023	10022		66	Female	29.5	Chinese	Never-Smoker	0
10024								

R interface

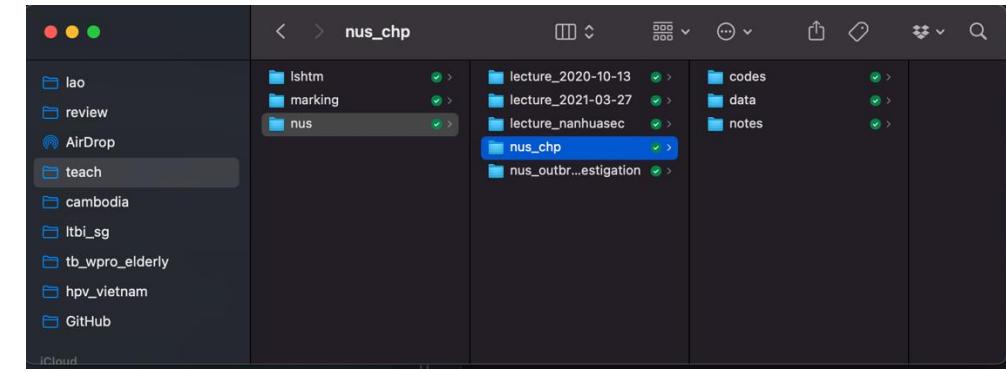
To use R, you need to have current versions of the following installed on your computer:

1. the [R software](#) itself, and
2. [RStudio Desktop](#) (we will only need to open this software, but RStudio needs R installed to work)

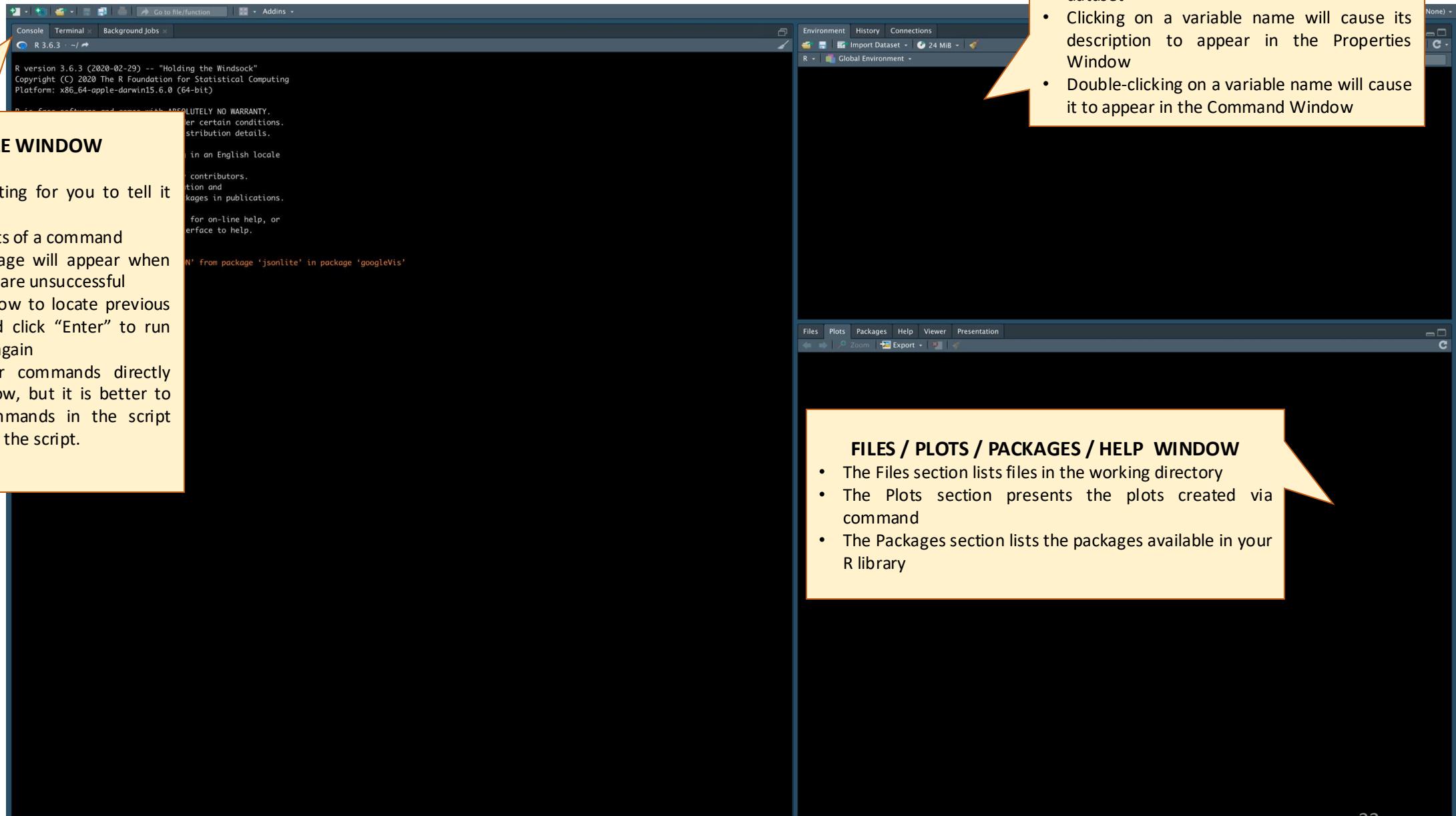


Before we start coding, can you locate:

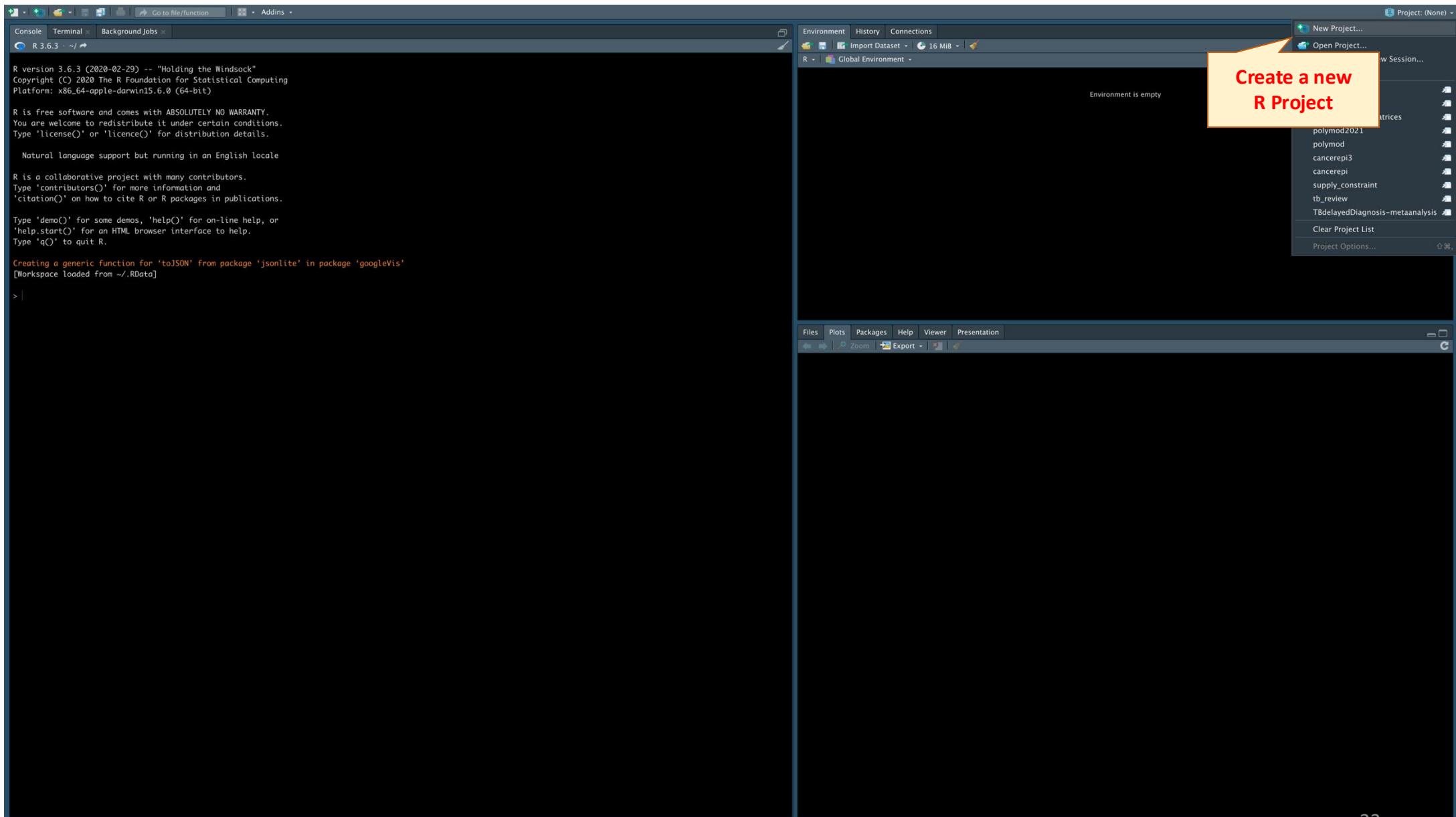
1. These **software packages (R and RStudio)**, and
2. The **folder with the dataset** – this will be your **working directory**



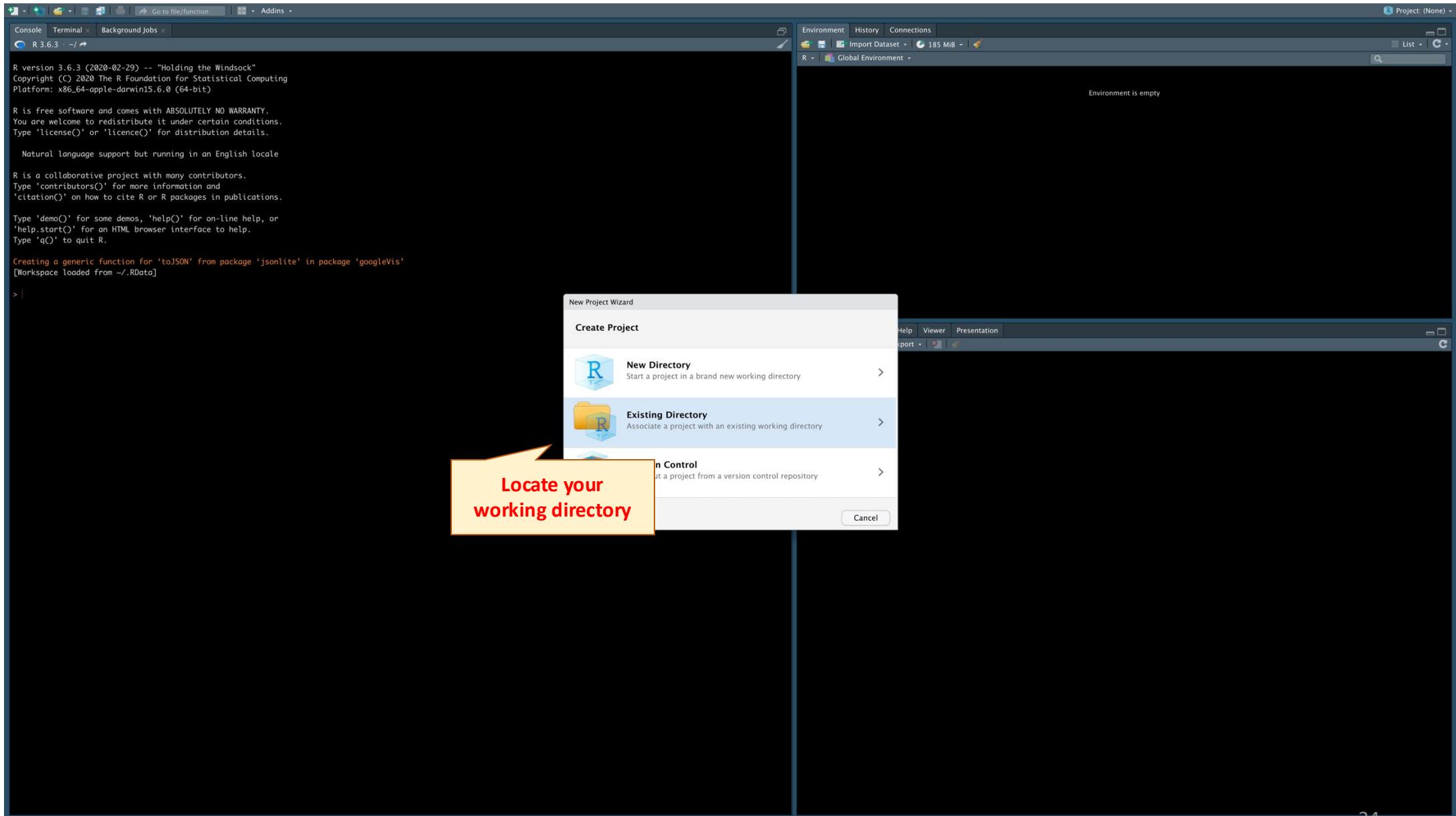
R interface



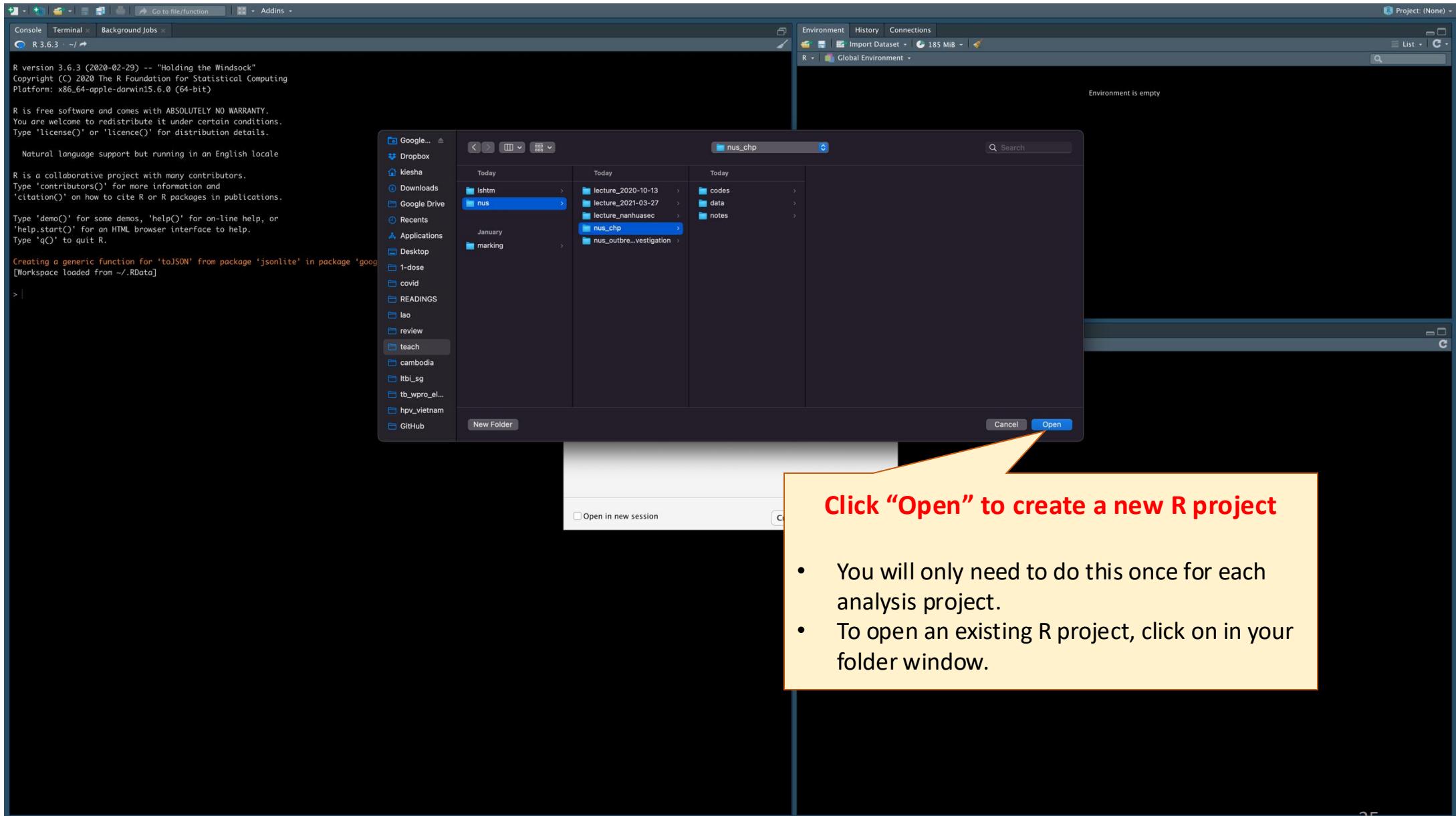
R interface – R Project (optional, but strongly recommended)



R interface – R Project (optional, but strongly recommended)



R interface – R Project (optional, but strongly recommended)



Click “Open” to create a new R project

- You will only need to do this once for each analysis project.
- To open an existing R project, click on in your folder window.

R interface – use **scripts** to write and save codes

The screenshot shows the RStudio interface. On the left, a yellow callout box highlights the 'Create a new R Script' option in the top menu bar. The main workspace shows a partial R script with comments about the R language and its distribution. Below the script, the R console shows a message about a function being created from a package, followed by a prompt '> |'. The right side of the interface shows the Environment pane with the message 'Environment is empty'.

Create a new R Script

- It is good coding practice to enter the commands in the script editor.
- Save your script in your working directory.

```
    "Using the Windsock"
  "for Statistical Computing"
  "64-bit)"

  ABSOLUTELY NO WARRANTY.
  under certain conditions.
  distribution details.

  g in an English locale
  y contributors.
  ation and
  ckages in publications.

  ' for on-line help, or
  terface to help.

Creating a generic function for 'toJSON' from package 'jsonlite' in package 'googleVis'
[Workspace loaded from ~/RData]

> |
```

Environment is empty

R interface – use scripts to write and save codes

The screenshot shows the RStudio interface. On the left, the R Script editor pane displays the R startup message and a command prompt (>). A yellow callout box with a red border and orange arrow points from the top-left towards the script editor, containing the text "Create a new R Script". The Global Environment pane on the right shows an empty environment with the message "Environment is empty". The top navigation bar includes tabs for Environment, History, and Connections, along with project and file management options.

Create a new R Script

- It is good coding practice to enter the commands in the script editor.
- Save your script in your working directory.

```
R version 3.6.3 (2020-02-29) -- "Holding the Windsock"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

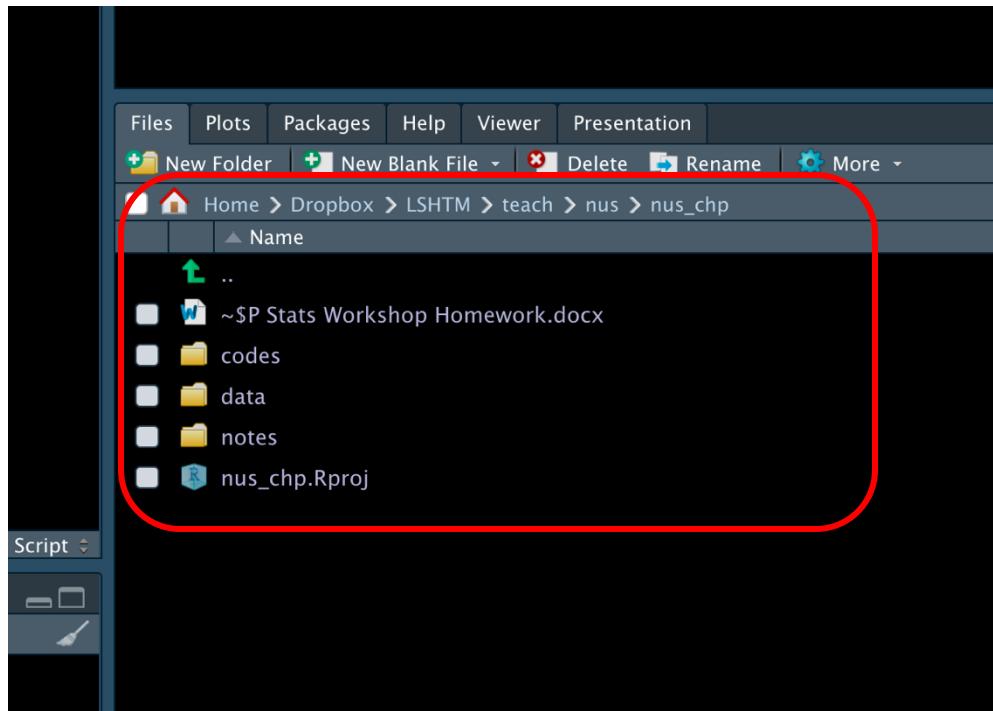
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

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.

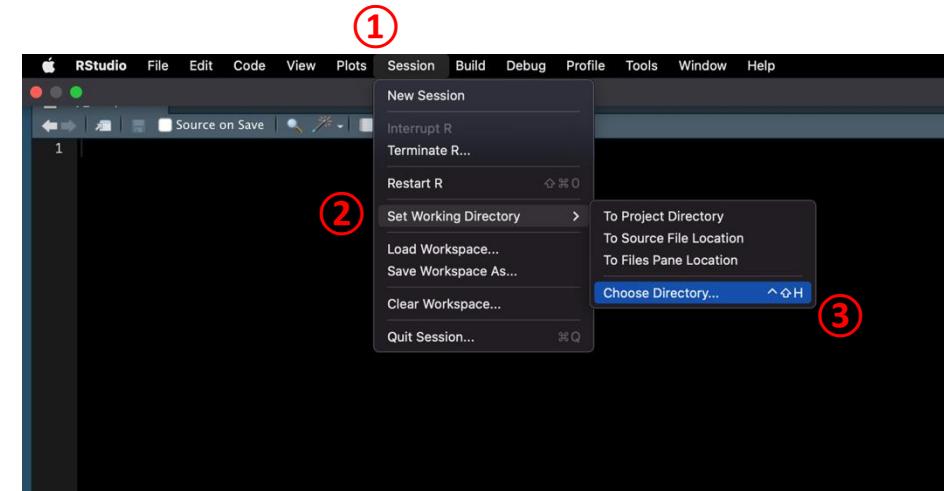
Creating a generic function for 'toJSON' from package 'jsonlite' in package 'googleVis'
[Workspace loaded from ~/RData]

> |
```

R project will load your working directory



But you can also use point-and-click menus...



- ① Select “Session”
- ② Select “Set Working Directory”
- ③ Browse and select the relevant folder

Checking your working directory:
`getwd()`

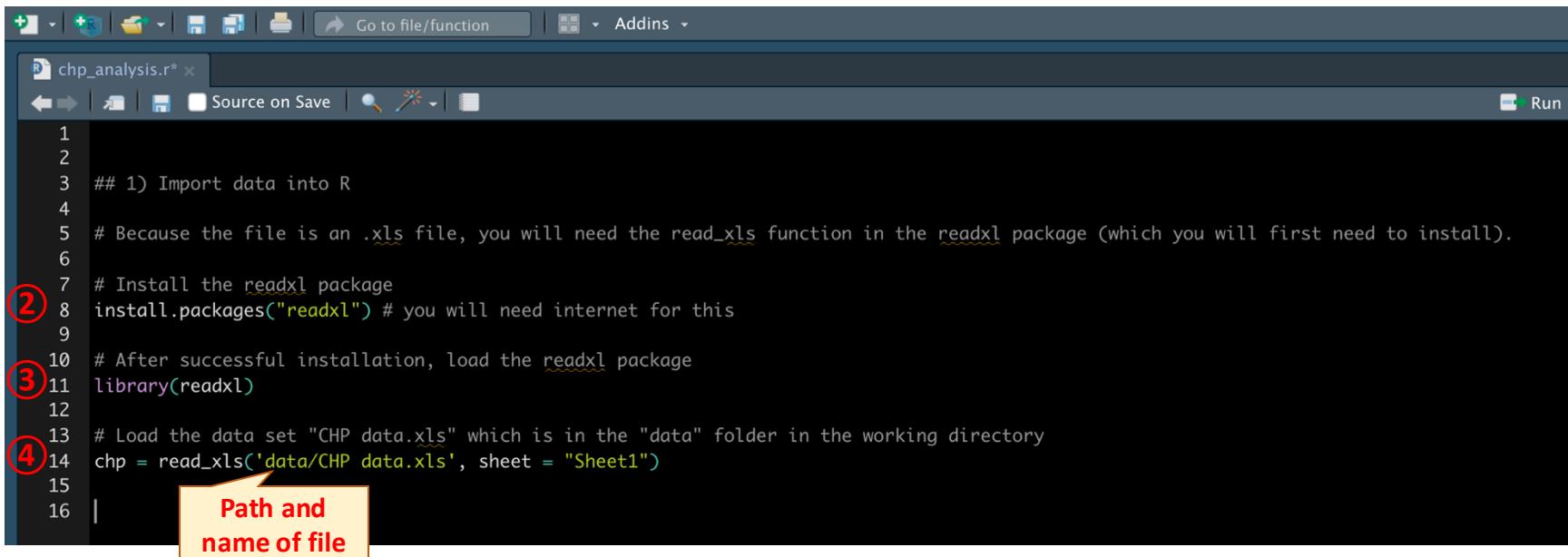
To run a command, click Run

Or press

- CTRL-Enter, on a Windows OS
- command-return on a Mac OS

Loading data files into R

- ① Locate the File on your computer, and identify the file type
- ② Because it is an .xls file, you will need to install a new R package – please make sure you are connected to the Internet
- ③ Load a new package
- ④ Load the data, identify the path to the file



```
1
2
3 ## 1) Import data into R
4
5 # Because the file is an .xls file, you will need the read_xls function in the readxl package (which you will first need to install).
6
7 # Install the readxl package
8 install.packages("readxl") # you will need internet for this
9
10 # After successful installation, load the readxl package
11 library(readxl)
12
13 # Load the data set "CHP data.xls" which is in the "data" folder in the working directory
14 chp = read_xls('data/CHP data.xls', sheet = "Sheet1")
15
16 |
```

Path and name of file

Commenting

- Use # signs to comment.
- *Comment liberally* in your R scripts.
- Anything to the right of a # is ignored by R.



SAVE
EARLY

SAVE
OFTEN

Loading data files into R

- ⑤ Check if the proper worksheet and cell range has been selected, click on the data set in the Environment window
- ⑥ Check the data

The screenshot shows the RStudio interface with two main windows open:

- Environment pane:** Shows the dataset "chp" imported into the "Global Environment". It contains 10022 observations and 8 variables. A red callout box with the text "⑤ Click on the data to view it" points to the dataset entry.
- Viewer pane:** Shows the "View" mode of the dataset "chp". The data is presented as a table with columns: id, age, gender, bmi, ethnicity, smoke, cvd, ldl, and s. The first few rows of data are visible.

Text annotations:

- ⑥ This is how the dataset looks like in the "View" mode**: A red callout box pointing to the viewer pane.
- Once the dataset is imported it will appear in the Environment window.**: A text box in the environment pane.

Console pane (bottom left):

```
R 3.6.3 (2020-02-29) -- "Holding the Windsock"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

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.

> library(readxl)
> chp = read_xlsx("data/CHP data.xls", sheet = "Sheet1")
Error: Evaluation error: zip file 'data/CHP data.xls' cannot be opened.
> ?read_xlsx
> chp = read_xls("data/CHP data.xls", sheet = "Sheet1")

> View(chp)
> View(chp)
> |
```

File Explorer pane (bottom right):

- Shows the file path: Home > Dropbox > LSHTM > teach > nus > nus_chp > data
- Lists the file "CHP data.xls" with a size of 1.5 MB and a modified date of Aug 1, 2022, 9:43 AM.

Now, let's have a look at the data in R!

We will follow the below colour scheme throughout the presentation!!



Describing the data

```
# Quick look at the data
head(chp)
str(chp)
summary(chp)
```



```
Console Terminal x Background Jobs x
R 3.6.3 · ~/Dropbox/LSHTM/teach/nus/nus_chp/ ↗
> # Quick look at the data
① > head(chp)
# A tibble: 6 × 8
  id    age gender   bmi ethnicity smoke      cvd   ldl
  <dbl> <dbl> <chr> <dbl> <chr> <chr> <dbl> <dbl>
1 1     72   Male   23.9 Indians Never-Smoker 0     3.49
2 2     73   Male   26.2 Chinese Never-Smoker 0     3.55
3 3     67   Male   19.9 Malays  Never-Smoker 0     3.15
4 4     65   Male   27.8 Indians Never-Smoker 0     2.97
5 5     72   Female 22   Indians Daily smoker 0     3.9
6 6     55   Male   20.9 Indians Never-Smoker 0     2.29

② > str(chp)
tibble [10,022 × 8] (S3:tbl_df/tbl/data.frame)
$ id     : num [1:10022] 1 2 3 4 5 6 7 8 9 10 ...
$ age    : num [1:10022] 72 73 67 65 72 55 72 66 66 62 ...
$ gender  : chr [1:10022] "Male" "Male" "Male" "Male" ...
$ bmi    : num [1:10022] 23.9 26.2 19.9 27.8 22 20.9 21.8 28.3 27.5 21.9 ...
$ ethnicity: chr [1:10022] "Indians" "Chinese" "Malays" "Indians" ...
$ smoke   : chr [1:10022] "Never-Smoker" "Never-Smoker" "Never-Smoker" "Never-Smoker" ...
$ cvd    : num [1:10022] 0 0 0 0 0 1 0 0 0 ...
$ ldl    : num [1:10022] 3.49 3.55 3.15 2.97 3.9 2.29 3.92 3.06 3.06 3.14 ...

③ > summary(chp)
   id          age        gender       bmi      ethnicity      smoke      cvd      ldl
Min.   : 1   Min.   : 2   Length:10022   Min.   :15.10  Length:10022   Length:10022   Min.   :0.0000  Min.   :-2.340
1st Qu.: 2506 1st Qu.: 63   Class :character 1st Qu.:22.10  Class :character  Class :character 1st Qu.:0.0000  1st Qu.: 2.940
Median : 5012 Median : 68   Mode  :character Median :25.20  Mode  :character  Mode  :character Median :0.0000  Median : 3.290
Mean   : 5012 Mean   : 68   NA's   :1           Mean   :25.38  NA's   :1           Mean   :0.1205  Mean   : 3.296
3rd Qu.: 7517 3rd Qu.: 73   NA's   :1           3rd Qu.:28.50  NA's   :1           3rd Qu.:0.0000  3rd Qu.: 3.650
Max.   :10022 Max.   :150  NA's   :1           Max.   :62.10  NA's   :1           Max.   :1.0000  Max.   :11.800
NA's   :4          NA's   :1
```

- ① the “head” function prints the top 6 rows of the data
 ② the “str” function provides a compact report on all variables in the dataset
 ③ the “summary” function provides a brief summary of the variables

Your task (~2 mins)
 • Review the output

Describing the data: review the output

- Numeric variables stored as byte, integer, long, float or double.
- Double type can hold non-integer numbers (i.e., decimals)

```
Console Terminal × Background Jobs ×
R 3.6.3 · ~/DropBox/LSHTM/teach/nus/nus_chp/ ↗
> # Quick look at the data
> head(chp)
# A tibble: 6 x 8
  id    age gender   bmi ethnicity smoke      cvd   ldl
  <dbl> <dbl> <chr> <dbl> <chr> <chr> <dbl> <dbl>
1 1     72   Male   23.9 Indians Never-Smoker 0     3.49
2 2     73   Male   26.2 Chinese
3 3     67   Male   19.9 Malays
4 4     65   Male   27.8 Indians
5 5     72   Female 22   Indians
6 6     55   Male   20.9 Indians Never-Smoker 2.29
> str(chp)
tibble [10,022 x 8] (S3:tbl_df/tbl/data.frame)
$ id    : num [1:10022] 1 2 3 4 5 6 7 8 9 10 ...
$ age   : num [1:10022] 72 73 67 65 72 55 72 66 66 62 ...
$ gender: chr [1:10022] "Male" "Male" "Male" "Male" ...
$ bmi   : num [1:10022] 23.9 26.2 19.9 27.8 22 20.9 21.8 28.3 27.5 21.9 ...
$ ethnicity: chr [1:10022] "Indians" "Chinese" "Malays" "Indians" ...
$ smoke  : chr [1:10022] "Never-Smoker" "Never-Smoker" "Never-Smoker" "Never-Smoker" ...
$ cvd    : num [1:10022] 0 0 0 0 0 1 0 0 0 ...
$ ldl    : num [1:10022] 3.49 3.55 3.15 2.97 3.9 2.29 3.92 3.06 3.06 3.14 ...
> summary(chp)
   id          age        gender       bmi      ethnicity      smoke      cvd      ldl
Min. : 1  Min. : 2  Length:10022  Min. :15.10  Length:10022  Length:10022  Min. :0.0000  Min. :-2.340
1st Qu.: 2506 1st Qu.: 63  Class :character 1st Qu.:22.10  Class :character  Class :character  1st Qu.:0.0000  1st Qu.: 2.940
Median : 5012 Median : 68  Mode  :character  Median :25.20  Mode  :character  Mode  :character  Median :0.0000  Median : 3.290
Mean   : 5012 Mean   : 68                    Mean   :25.38                    Mean   :0.1205  Mean   : 3.296
3rd Qu.: 7517 3rd Qu.: 73                    3rd Qu.:28.50                    3rd Qu.:0.0000  3rd Qu.: 3.650
Max.   :10022 Max.   :150                   Max.  :62.10                   Max.  :1.0000  Max.  :11.800
NA's   : 4                           NA's   :1                           NA's   :2
```

- Number of participants (= 10 022)
- Number of variables (= 8)

Character/factor/string variables contain non-numeric characters in them, including symbols like commas, hyphens and periods.

You need to be extra careful when doing regression analysis with a character/factor variable in R.

Describing the data: review the output

```

Console Terminal × Background Jobs ×
R 3.6.3 · ~/Dropbox/LSHTM/teach/nus/nus_chp/ ↗
> # Quick look at the data
> head(chp)
# A tibble: 6 x 8
  id    age gender   bmi ethnicity smoke      cvd    ldl
  <dbl> <dbl> <chr> <dbl> <chr> <chr> <dbl> <dbl>
1    1     72 Male    23.9 Indians Never-Smoker 0    3.49
2    2     73 Male    26.2 Chinese Never-Smoker 0    3.55
3    3     67 Male    19.9 Malays  Never-Smoker 0    3.15
4    4     65 Male    27.8 Indians Never-Smoker 0    2.97
5    5     72 Female  22   Indians Daily smoker 0    3.9
6    6     55 Male    20.9 Indians Never-Smoker 0    2.29
> str(chp)
tibble [10,022 x 8] (S3:tbl_df/tbl/data.frame)
$ id    : num [1:10022] 1 2 3 4 5 6 7 8 9 10 ...
$ age   : num [1:10022] 72 73 67 65 72 55 72 66 66 62 ...
$ gender: chr [1:10022] "Male" "Male" "Male" "Male" ...
$ bmi   : num [1:10022] 23.9 26.2 19.9 27.8 22 20.9 21.8 28.3 27.5 21.9 ...
$ ethnicity: chr [1:10022] "Indians" "Chinese" "Malays" "Indians" ...
$ smoke  : chr [1:10022] "Never-Smoker" "Never-Smoker" "Never-Smoker" "Never-Smoker" ...
$ cvd    : num [1:10022] 0 0 0 0 0 0 1 0 0 0 ...
$ ldl    : num [1:10022] 3.49 3.55 3.15 2.97 3.9 2.29 3.92 3.06 3.06 3.14 ...
> summary(chp)
   id          age        gender       bmi      ethnicity      smoke      cvd      ldl
  : 1  Min.   : 2  Length:10022  Min.   :15.10  Length:10022  Length:10022  Min.   :0.0000  Min.   :-2.340
Qu.: 2506  1st Qu.: 63  Class :character  1st Qu.:22.10  Class :character  Class :character  1st Qu.:0.0000  1st Qu.: 2.940
  : 5012  Median : 68  Mode  :character  Median :25.00  Mode  :character  Mode  :character  Median :0.0000  Median : 3.290
  : 5012  Mean   : 68  Mode  :character  Mean   :25.38  Mode  :character  Mode  :character  Mean   :0.1205  Mean   : 3.296
  : 7517  3rd Qu.: 73  Mode  :character  3rd Qu.:28.50  Mode  :character  Mode  :character  3rd Qu.:0.0000  3rd Qu.: 3.650
  :10022  Max.   :150  Mode  :character  Max.   :62.10  Mode  :character  Mode  :character  Max.   :1.0000  Max.   :11.800
   NA's   :4           NA's   :1           NA's   :2

```

All participants have unique IDs

Data appears to be missing for some participants

Complete data present for "cvd" with 2 unique categories:
Yes = 1
No = 0

Some values seem out of range

LDL values cannot be negative

This LDL value appears to be too high

Describing the data: review the output

What about the character variables?

```
table(chp$gender, useNA = 'ifany')
table(chp$ethnicity, useNA = 'ifany')
table(chp$smoke, useNA = 'ifany')
```

<<DATA NAME>>\$<<VARIABLE NAME>>

```
Console Terminal × Background Jobs ×
R 3.6.3 · ~/Dropbox/LSHTM/teach/nus/nus_chp/ ↗
> # Quick look at the data
> head(chp)
# A tibble: 6 x 8
  id    age gender   bmi ethnicity smoke      cvd   ldl
<dbl> <dbl> <chr> <dbl> <chr> <chr> <dbl> <dbl>
1 1     72   Male   23.9 Indians Never-Smoker 0   3.49
2 2     73   Male   26.2 Chinese Never-Smoker 0   3.55
3 3     67   Male   19.9 Malays  Never-Smoker 0   3.15
4 4     65   Male   27.8 Indians Never-Smoker 0   2.97
5 5     72   Female 22   Indians Daily smoker 0   3.9
6 6     55   Male   20.9 Indians Never-Smoker 0   2.29
> str(chp)
tibble [10,022 x 8] (S3:tbl_df/tbl/data.frame)
$ id     : num [1:10022] 1 2 3 4 5 6 7 8 9 10 ...
$ age    : num [1:10022] 72 73 67 65 72 55 72 66 66 62 ...
$ gender  : chr [1:10022] "Male" "Male" "Male" "Male" ...
$ bmi    : num [1:10022] 23.9 26.2 19.9 27.8 22 20.9 21.8 28.3 27.5 21.9 ...
$ ethnicity: chr [1:10022] "Indians" "Chinese" "Malays" "Indians" ...
$ smoke   : chr [1:10022] "Never-Smoker" "Never-Smoker" "Never-Smoker" "Never-Smoker" ...
$ cvd     : num [1:10022] 0 0 0 0 0 0 1 0 0 0 ...
$ ldl     : num [1:10022] 3.49 3.55 3.15 2.97 3.9 2.29 3.92 3.06 3.06 3.14 ...
> summary(chp)
   id        age       gender      bmi   ethnicity      smoke      cvd      ldl
Min. : 1  Min. : 2  Length:10022  Min. :15.10  Length:10022  Length:10022  Min. :0.0000  Min. :-2.340
1st Qu.: 2506  1st Qu.: 63  Class :character  1st Qu.:22.10  Class :character  Class :character  1st Qu.:0.0000  1st Qu.: 2.940
Median : 5012  Median : 68  Mode  :character  Median :25.20  Mode  :character  Mode  :character  Median :0.0000  Median : 3.290
Mean   : 5012  Mean   : 68  NA's   :1          Mean   :25.38  NA's   :1          Mean   :0.1205  Mean   : 3.296
3rd Qu.: 7517  3rd Qu.: 73  NA's   :1          3rd Qu.:28.50  NA's   :1          3rd Qu.:0.0000  3rd Qu.: 3.650
Max.   :10022  Max.   :150  NA's   :4          Max.   :62.10  NA's   :1          Max.   :1.0000  Max.   :11.800
NA's   :2
```

Describing the data: review the output

What about the character variables?

```
> # Exploring the data: character variables  
> table(chp$gender, useNA = 'ifany')
```

Female	male	Male	q	<NA>
5028	1	4989	2	2

Gender should have 2 unique values: Male and Female

Data appears to be missing for some participants

```
> table(chp$ethnicity, useNA = 'ifany')
```

chinese	Chinese	Indians	Malays	<NA>
1	3882	3087	3049	3

Ethnicity should have 3 unique values

```
> table(chp$smoke, useNA = 'ifany')
```

Daily smoker	Ex-smoker	Never-Smoker	Occasional smoker	zsmoker	<NA>
1480	1120	6517	901	2	2

Smoke should have 4 unique values

Summarising the data (continuous variables)

The screenshot shows the RStudio interface. In the top-left pane, there is an R script named "chp_analysis.r". The code includes:

```

22 summary(chp)
23
24 # Exploring the data: character variables
25 table(chp$gender, useNA = 'ifany')
26 table(chp$ethnicity, useNA = 'ifany')
27 table(chp$smoke, useNA = 'ifany')
28
29
30 # Summarize the data: continuous variable
31 summary(chp$ldl)
32 quantile(chp$ldl)
33 quantile(chp$ldl,na.rm = TRUE)
34 hist(chp$ldl,breaks = 100,xlab="LDL Cholesterol (mmol/L)", main="Histogram of LDL Cholesterol")
35
36
37
38
39
40 summary(chp$bmi)
41 table(chp$ethnicity, useNA = 'ifany')
42 table(chp$smoke, useNA = 'ifany')
43
44
45 prop.table(table(chp$gender))
46

```

In the bottom-left pane, the R console shows the output of the commands:

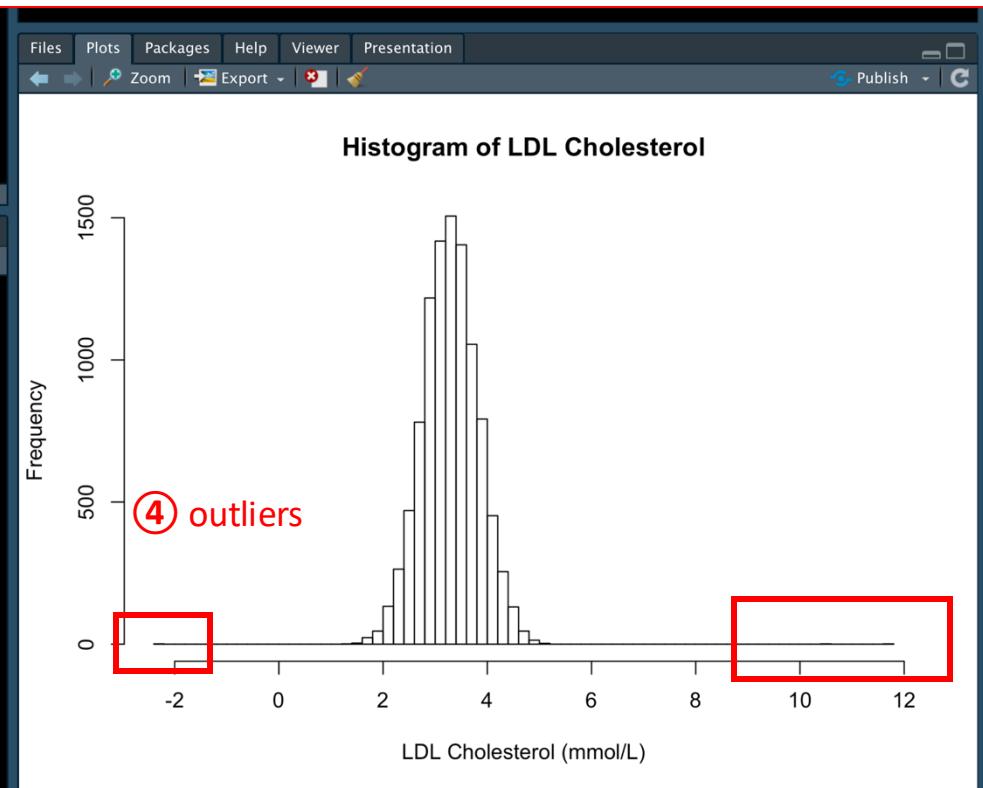
```

> # Summarize the data: continuous variable
> summary(chp$ldl)
   Min. 1st Qu. Median 3rd Qu. Max. NA's
-2.340  2.940  3.290  3.296  3.650 11.800    2 ①
> quantile(chp$ldl)
Error in quantile.default(chp$ldl) :
  missing values and NaN's not allowed if 'na.rm' is FALSE
> quantile(chp$ldl,na.rm = TRUE)
  0% 25% 50% 75% 100%
-2.34  2.94  3.29  3.65 11.80 ②
> hist(chp$ldl,breaks = 100,xlab="LDL Cholesterol (mmol/L)", main="Histogram of LDL Cholesterol") ③
>

```

Red circles numbered 1 through 4 point to specific parts of the code and output.

- ① Use the “summary” function to summarise a continuous variable
- ② The “quantile” function produces sample quantiles corresponding to the given probabilities. The default probabilities are 0, 0.25, 0.5, 0.75, and 1; they represent the minimum, first quartile, median, third quartile, and maximum values, respectively.
- ③ Plot the histogram of the variable
- ④ Study the histogram, can you identify any outliers? (Hint: look very carefully)



Summarising the data (continuous variables)

chp_analysis.r*

```

22 summary(cnp)
23
24 # Exploring the data: character variables
25 table(chp$gender, useNA = 'ifany')
26 table(chp$ethnicity, useNA = 'ifany')
27 table(chp$smoke, useNA = 'ifany')
28
29
30 # Summarize the data: continuous variable
31 summary(chp$ldl)
32 quantile(chp$ldl)
33 quantile(chp$ldl,na.rm = TRUE)
34 hist(chp$ldl,breaks = 100,xlab="LDL Cholesterol (mmol/L)", main="Histogram of LDL Cholesterol")
35
36
37
38
39
40 summary(chp$bmi)
41 table(chp$ethnicity, useNA = 'ifany')
42 table(chp$smoke, useNA = 'ifany')
43
44
45 prop.table(table(chp$gender))
46

```

Console

Erroneous data : negative values

```

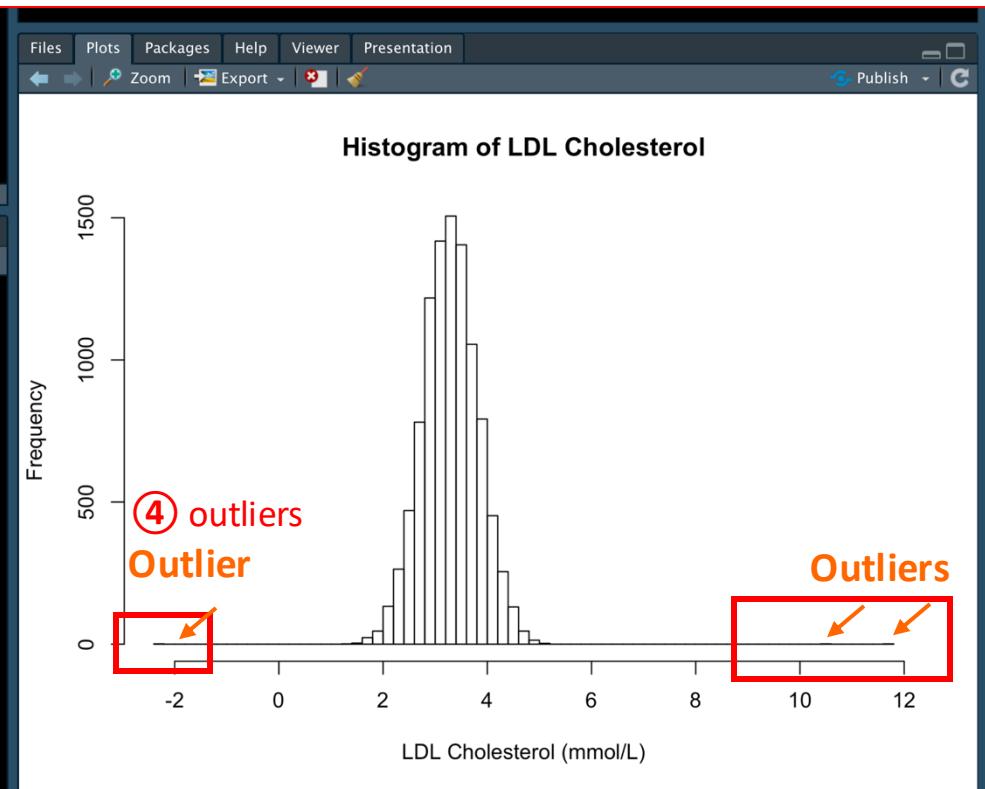
R 3.6.3 - ~/Desktop/nus/nus_chp/
> # Summarizing data: continuous variable
> summary(chp$ldl)
  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
-2.340 2.940 3.290 3.296 3.650 11.800 2
> quantile(chp$ldl)
Error in quantile.default(chp$ldl) :
  missing values and NaN's not allowed if 'na.rm' is FALSE
> quantile(chp$ldl,na.rm = TRUE)
  0% 25% 50% 75% 100%
-2.34 2.94 3.29 3.65 11.80
> hist(chp$ldl,breaks = 100,xlab="LDL Cholesterol (mmol/L)", main="Histogram of LDL Cholesterol")
>

```

Some values appear to be too high

LDL values missing for 2 participants

- ① Use the “summary” function to summarise a continuous variable
- ② The “quantile” function produces sample quantiles corresponding to the given probabilities. The default probabilities are 0, 0.25, 0.5, 0.75, and 1; they represent the minimum, first quartile, median, third quartile, and maximum values, respectively.
- ③ Plot the histogram of the variable
- ④ Study the histogram, can you identify any outliers? (Hint: look very carefully)



Summarising the data (continuous variables)

Plot the box plot

```
boxplot(chp$ldl, ylab="LDL Cholesterol (mmol/L)", main="Boxplot of LDL Cholesterol", col = 'grey')
```

Boxplot of LDL Cholesterol

The boxplot displays the distribution of LDL cholesterol levels. The y-axis is labeled "LDL Cholesterol (mmol/L)" and ranges from -2 to 12. The x-axis is labeled "LDL Cholesterol (mmol/L)" and ranges from -2 to 12. The box represents the interquartile range (IQR) from approximately 2.5 to 5.5 mmol/L, with a median line at about 3.5 mmol/L. Whiskers extend from the box to approximately 1.5 and 5.5 mmol/L. Two outliers are identified with red circles and labeled "④ outliers": one at approximately 11.5 mmol/L and another at approximately 10.5 mmol/L.

④ outliers

LDL Cholesterol (mmol/L)

LDL Cholesterol (mmol/L)

39

In-class Exercise #1

Summarize the data for the following continuous variables:

- (i) Age
- (ii) BMI

Do you observe any missing/erroneous values?

In-class Exercise #1 Solution

```

> # Age
> summary(chp$age)
  Min.   1st Qu.    Median      Mean   3rd Qu. 
  2       63       68       68       73 
  NA's:4
> quantile(chp$age,na.rm = TRUE)
  0%  25%  50%  75% 100% 
  2   63   68   73  150 
> # Checking the distribution of the variable (using histogram)
> hist(chp$age,breaks = 100,xlab="Age (in years)", main="Histogram of Age")
> # Checking for outliers (using box plot)
> boxplot(chp$age,ylab="Age (in years)", main="Boxplot of Age",col = 'grey')
>

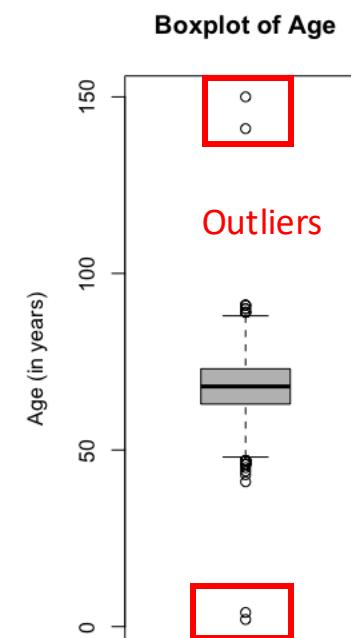
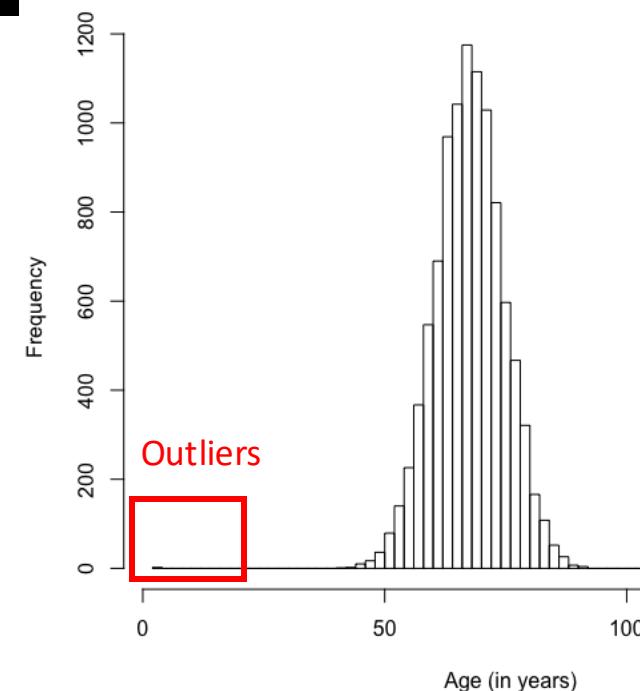
```

Values out of range

Max.

150

Age missing for some participants



In-class Exercise #1 Solution

```

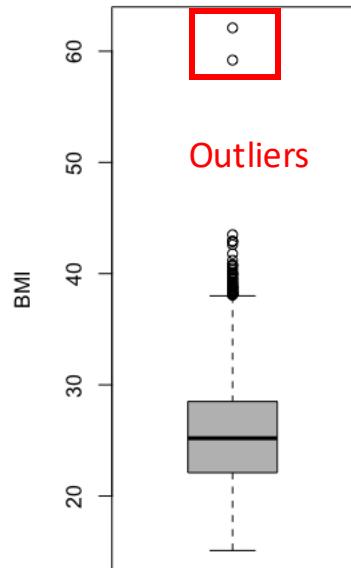
> # BMI
> summary(chp$bmi)
  Min. 1st Qu. Median Mean 3rd Qu.
  15.10 22.10 25.20 25.38 28.50
  Max. NA's
  62.10 1
> quantile(chp$bmi,na.rm = TRUE)
  0% 25% 50% 75% 100%
  15.1 22.1 25.2 28.5 62.1
> # Checking the distribution of the variable (using histogram)
> hist(chp$bmi,breaks = 100,xlab="BMI", main="Histogram of BMI")
> # Checking for outliers (using box plot)
> boxplot(chp$bmi,ylab="BMI", main="Boxplot of BMI",col = 'grey')
>

```

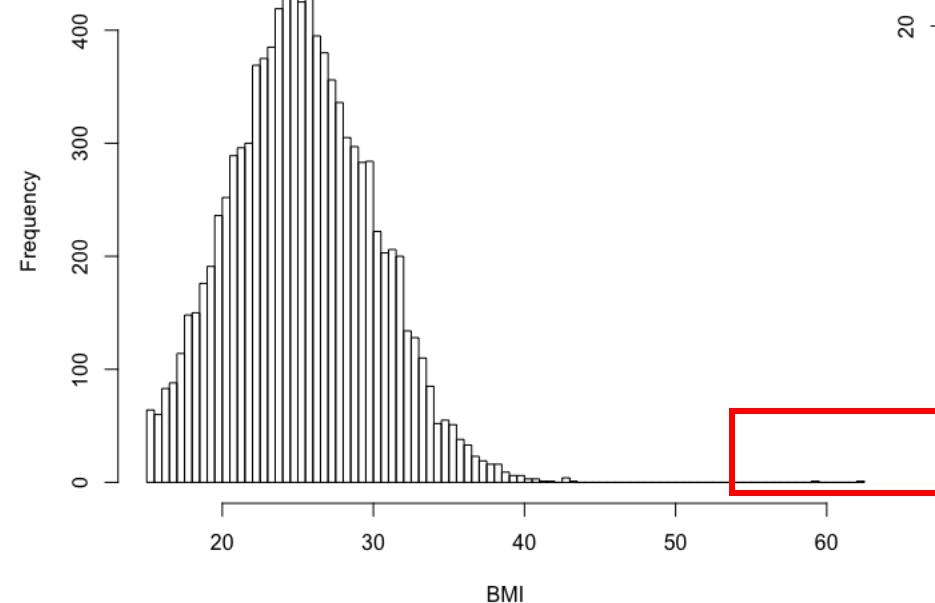
Some values appear
to be too high

BMI missing for some
participants

Boxplot of BMI



Histogram of BMI



Summarising the data (categorical variables)

- ① Use the “table” function to build a (one-way) contingency table of the counts at each combination of factor levels.
- ② Use the “prop.table” function to get the proportions.

```
> # Gender
① > table(chp$gender, useNA = 'ifany')

Female    male    Male      q    <NA>
  5028     1    4989     2      2

② > prop.table(table(chp$gender, useNA = 'ifany'))

    Female        male        Male        q        <NA>
0.50169626821 0.00009978048 0.49780482938 0.00019956097 0.00019956097
> |
```

Summarising the data (categorical variables)

- ① Use the “table” function to build a (one-way) contingency table of the counts at each combination of factor levels.
- ② Use the “prop.table” function to get the proportions.

```
> # Gender
① > table(chp$gender, useNA = 'ifany')

Female male Male q <NA>
5028 1 4989 2 2

Error: Gender marked as "q" for 2 participants

② > prop.table(table(chp$gender, useNA = 'ifany'))

Female male Male q <NA>
0.50169626821 0.00009978048 0.49780482938 0.00019956097 0.00019956097

"Male" spelled as "male" for 1 participant

Gender missing for 2 participants

③ Calculate % from proportions
```

In-class Exercise #2

Summarize the data for the following categorical variables:

- (i) Smoker
- (ii) Ethnicity

Do you observe any missing/erroneous values?

In-class Exercise #2 Solution

Summarising the data (categorical variables)

```
> # Smoke
> table(chp$smoke, useNA = 'ifany')
```

Daily smoker	Ex-smoker	Never-Smoker	Occasional smoker
1480	1120	6517	901

```
> prop.table(table(chp$smoke, useNA = 'ifany'))
```

Daily smoker	Ex-smoker	Never-Smoker	Occasional smoker	zsmoker	<NA>
0.147675115	0.111754141	0.650269407	0.089902215	0.000199561	0.000199561

```
>
```

```
> # Ethnicity
```

```
> table(chp$ethnicity, useNA = 'ifany')
```

chinese	Chinese	Indians	Malays	<NA>
1	3882	3087	3049	3

```
> prop.table(table(chp$ethnicity, useNA = 'ifany'))
```

chinese	Chinese	Indians	Malays	<NA>
0.00009978048	0.38734783476	0.30802235083	0.30423069248	0.00029934145

“Chinese” spelled as
“chinese” for 1 participant

Error: smoke marked as “zsmoker” for
2 participants

zsmoker
2

Smoking status
data missing for 2
participants

<NA>
2

Ethnicity data missing
for 3 participants

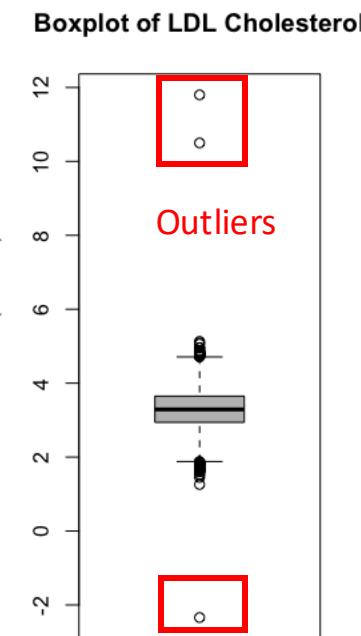
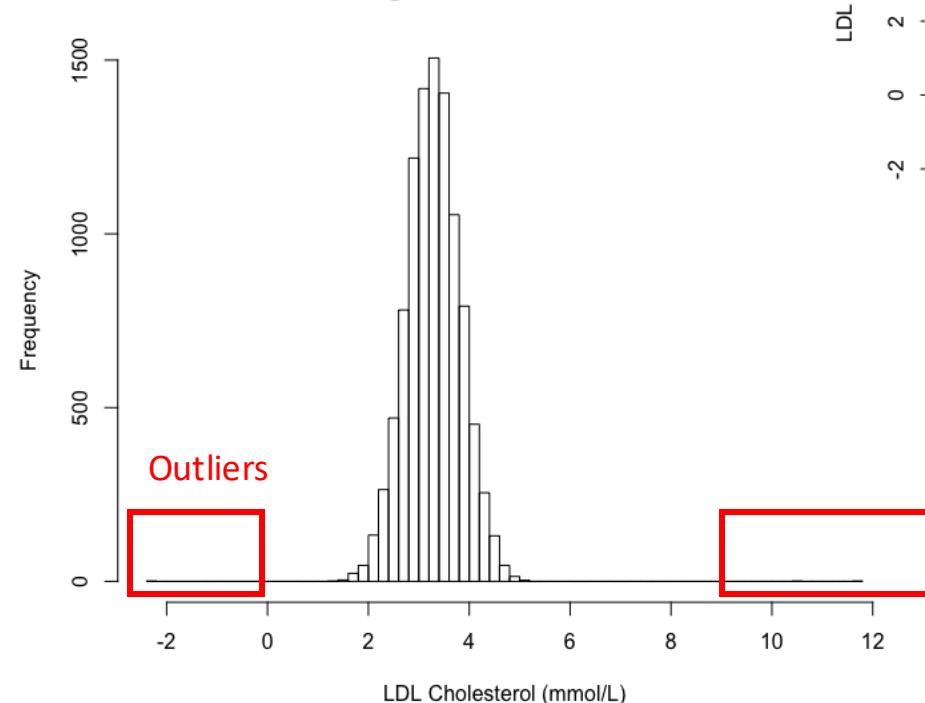
Data cleaning for continuous variable

Data cleaning for continuous variable: LDL Cholesterol (recap)

```
> summary(chp$ldl)
   Min. 1st Qu. Median    Mean 3rd Qu.   Max.   NA's
 -2.340  2.940  3.290  3.296  3.650  11.800      2
> quantile(chp$ldl,na.rm = TRUE)
  0% 25% 50% 75% 100%
-2.34 2.94 3.29 3.65 11.80
> # Checking the distribution of the variable (using histogram)
> hist(chp$ldl,breaks = 100,xlab="LDL Cholesterol (mmol/L)", main="Histogram of LDL Cholesterol")
> # Checking for outliers (using box plot)
> boxplot(chp$ldl,ylab="LDL Cholesterol (mmol/L)", main="Boxplot of LDL Cholesterol",col = 'grey')
> |
```

RECAP

Erroneous data : negative values
 LDL values missing for 2 participants
 Some values appear to be too high



Data cleaning for continuous variable: LDL Cholesterol (recap)

```
> summary(chp$ldl)
  Min. 1st Qu. Median   Mean 3rd Qu.   Max.   NA's
-2.340  2.940  3.290  3.296  3.650  11.800      2
> quantile(chp$ldl,na.rm = TRUE)
  0% 25% 50% 75% 100%
-2.34  2.94  3.29  3.65 11.80
> # Checking the distribution of the variable (using histogram)
> hist(chp$ldl,breaks = 100,xlab="LDL Cholesterol (mmol/L)", main="Histogram of LDL Cholesterol")
> # Checking for outliers (using box plot)
> boxplot(chp$ldl,ylab="LDL Cholesterol (mmol/L)", main="Boxplot of LDL Cholesterol",col = 'grey')
>
```

RECAP

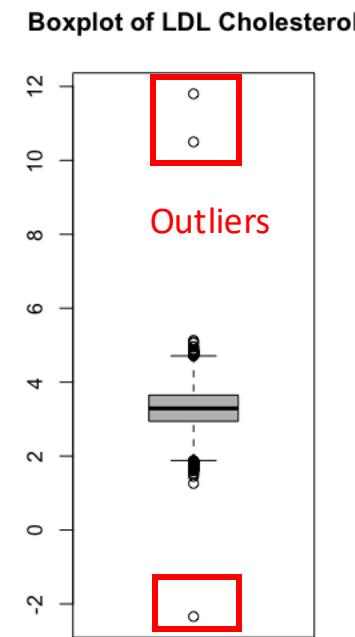
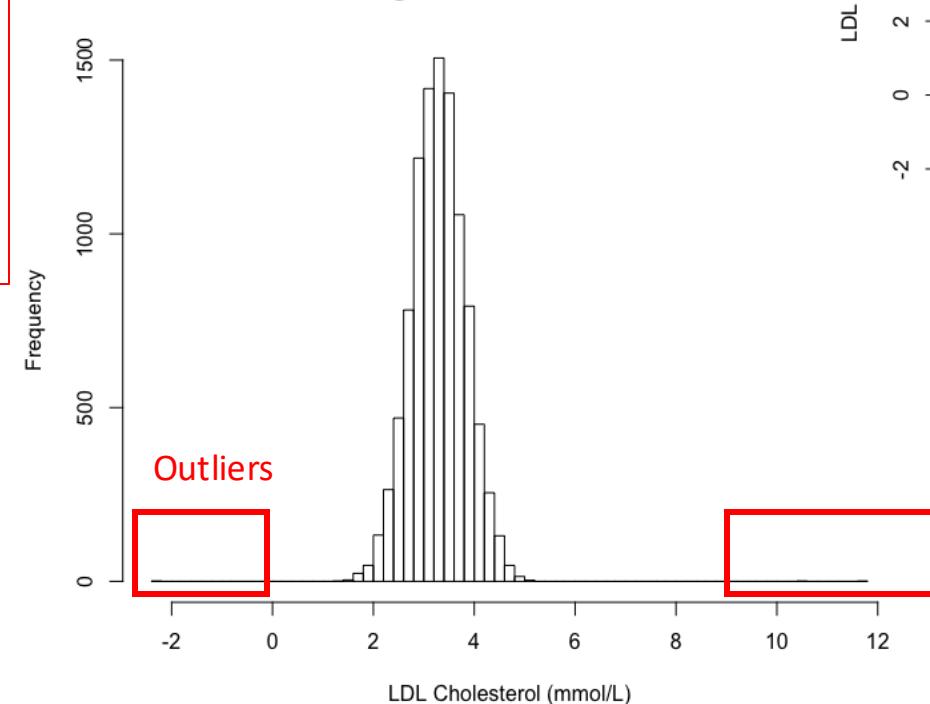
Erroneous data : negative values
 LDL values missing for 2 participants

Some values appear to be too high

- ① Create new variable “ldl_drop” which has the same values as ldl (for now).
- ② For “ldl_drop” discard those outside the range of 1-10
- ③ Summarise the new variable

```
① > # LDL
  > chp$ldl_drop = chp$ldl
  > chp$ldl_drop[chp$ldl_drop<1 | chp$ldl_drop>10]
  [1] -2.34 11.80 10.50    NA    NA
② > chp$ldl_drop[chp$ldl_drop<1 | chp$ldl_drop>10] = NA
③ > summary(chp$ldl_drop)
  Min. 1st Qu. Median   Mean 3rd Qu.   Max.   NA's
  1.260  2.940  3.290  3.295  3.650  5.130      5
```

Verify the range



In-class Exercise #3

Data cleaning for the following continuous variables:

- (i) Age
- (ii) BMI

In-class Exercise #3 Solution

```

> # Age
> summary(chp$age)
  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
  2       63     68   68    73   150      4
> quantile(chp$age,na.rm = TRUE)
  0% 25% 50% 75% 100%
  2   63   68   73   150
> # Checking the distribution of the variable (using histogram)
> hist(chp$age,breaks = 100,xlab="Age (in years)", main="Histogram of Age")
> # Checking for outliers (using box plot)
> boxplot(chp$age,ylab="Age (in years)", main="Boxplot of Age",col = 'grey')
>

```

Values out of range

Min.
2

Age missing for some participants

Max.
150

NA's
4

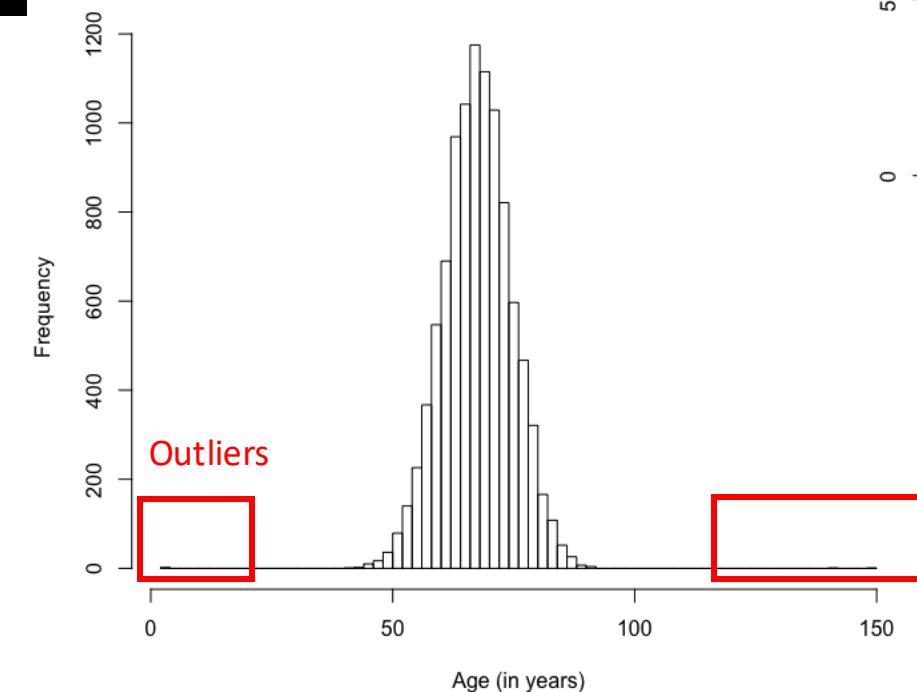
- ① Create new variable “age_drop” which has the same values as age (for now).
- ② For “age_drop” discard those outside the range of 40-100
- ③ Summarise the new variable

```

> # Age
① > chp$age_drop = chp$age
> chp$age_drop[chp$age_drop<40 | chp$age_drop>100]
[1] 4 141 150 2 NA NA NA NA
② > chp$age_drop[chp$age_drop<40 | chp$age_drop>100] = NA
③ > summary(chp$age_drop)
  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
  41.00 63.00 68.00 67.99 73.00 91.00      8

```

Verify the range



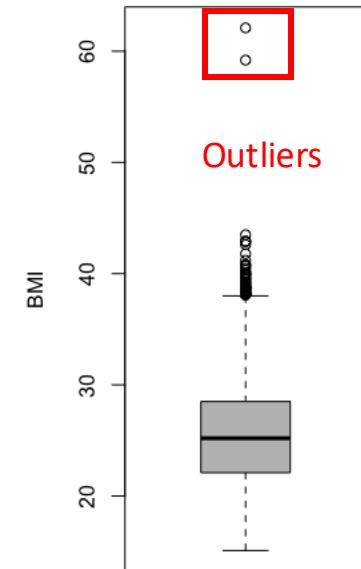
In-class Exercise #3 Solution

```
> # BMI
> summary(chp$bmi)
  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
  15.10 22.10 25.20 25.38 28.50 62.10 1
> quantile(chp$bmi,na.rm = TRUE)
  0% 25% 50% 75% 100%
15.1 22.1 25.2 28.5 62.1
> # Checking the distribution of the variable (using histogram)
> hist(chp$bmi,breaks = 100,xlab="BMI", main="Histogram of BMI")
> # Checking for outliers (using box plot)
> boxplot(chp$bmi,ylab="BMI", main="Boxplot of BMI",col = 'grey')
>
```

Some values appear
to be too high

BMI missing for some
participants

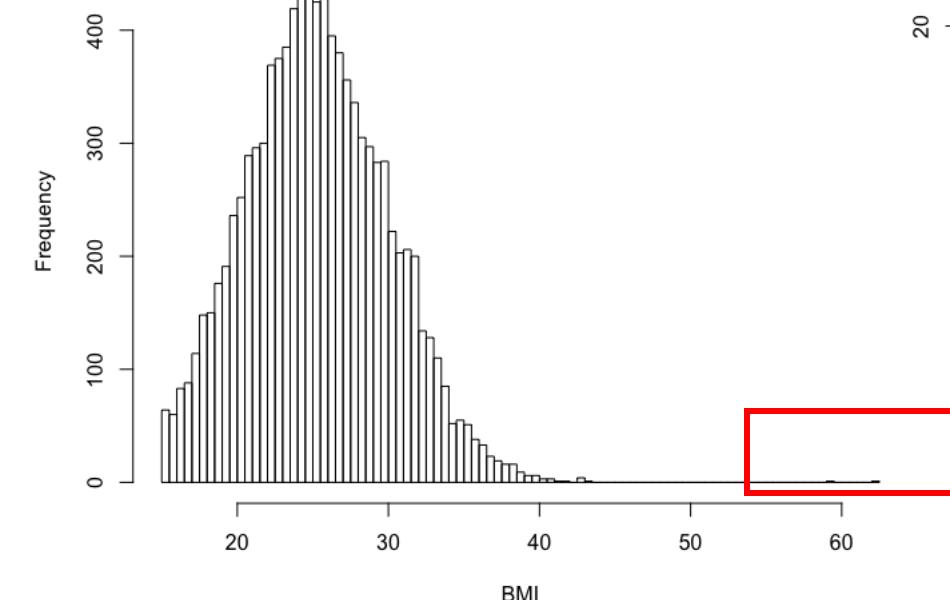
Boxplot of BMI



- ① Create new variable "bmi_drop" which has the same values as age (for now).
- ② For "bmi_drop" discard those > 50
- ③ Summarise the new variable

```
① > chp$bmi_drop = chp$bmi
> chp$bmi_drop[chp$bmi_drop>50]
[1] NA 62.1 59.2
② > chp$bmi_drop[chp$bmi_drop>50] = NA
③ > summary(chp$bmi_drop)
  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
  15.10 22.10 25.20 25.37 28.50 43.50 3
```

Verify the range



Data cleaning for categorical variable: gender (recap)

```
> # Gender
> table(chp$gender, useNA = 'ifany')

Female   male   Male    q   <NA>
5028     1     4989   2     2

Error: Gender marked as "q" for 2 participants

Gender missing for 2 participants

> prop.table(table(chp$gender, useNA = 'ifany'))

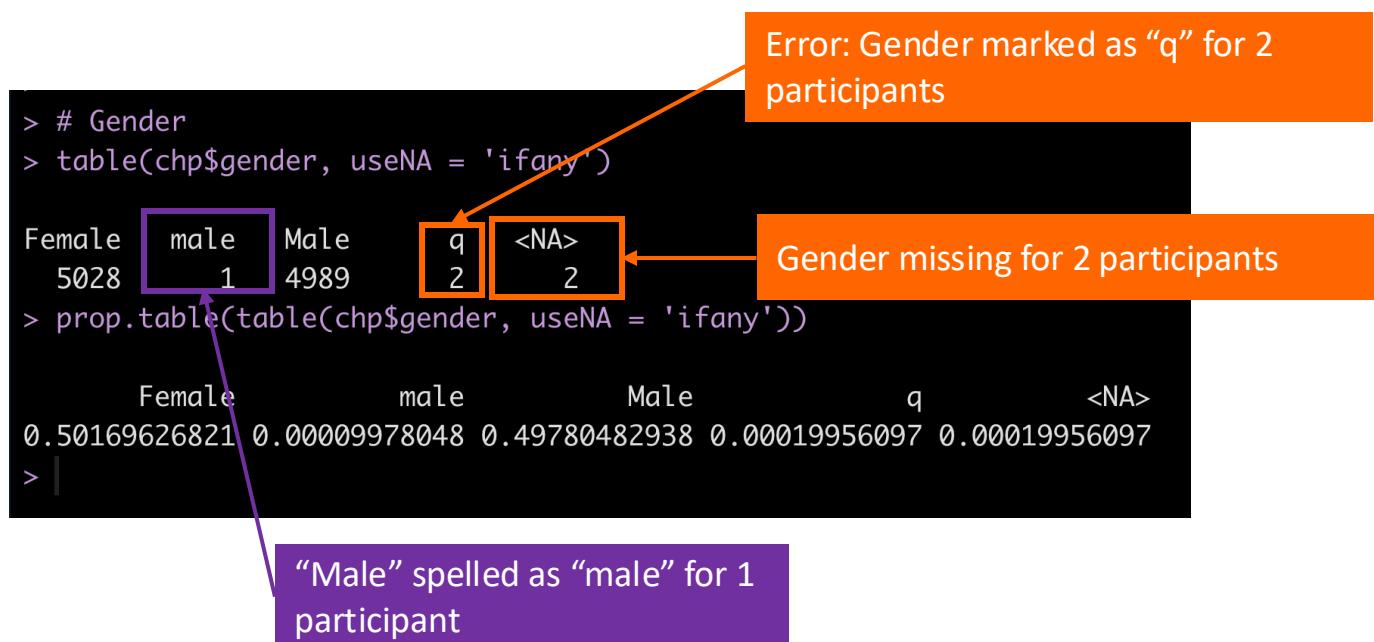
Female      male      Male       q      <NA>
0.50169626821 0.00009978048 0.49780482938 0.0019956097 0.0019956097

>

"Male" spelled as "male" for 1 participant
```

- ✓ Erroneous data; 2 observations have values “q”
- ✓ “Male” also spelled in lowercase “male”

Data cleaning for categorical variable: gender



- ✓ "Male" also spelled in lowercase "male"
- ✓ Erroneous data; 2 observations have values "q"

- ① Replace "male" with "Male" using the `%in%` operator
- ② Replace the errors (i.e., "q") as missing values using the `%in%` operator
- ③ Verify the changes

```

> # gender
> table(chp$gender, useNA = 'ifany')

Female   male   Male    q    <NA>
5028      1    4989    2    2
> chp$gender[chp$gender %in% "male"]
[1] "male"
> chp$gender[chp$gender %in% "male"] = "Male"
> chp$gender[chp$gender %in% "q"]
[1] "q" "q"
> chp$gender[chp$gender %in% "q"] = NA
> table(chp$gender, useNA = 'ifany')

```

Female	Male	<NA>
5028	4990	4

Verify the values

In-class Exercise #4

Data cleaning for the following categorical variable:

- (i) Smoke
- (ii) Ethnicity

In-class Exercise #4 Solution

Data cleaning for categorical variable

```
> # Smoke
> table(chp$smoke, useNA = 'ifany')
```

Daily smoker	Ex-smoker	Never-Smoker	Occasional smoker	
1480	1120	6517	901	

```
> chp$smoke[chp$smoke %in% "zsmoker"]
```

```
[1] "zsmoker" "zsmoker"
```

```
> chp$smoke[chp$smoke %in% "zsmoker"] = NA ①
```

```
> table(chp$smoke, useNA = 'ifany')
```

Daily smoker	Ex-smoker	Never-Smoker	Occasional smoker	
1480	1120	6517	901	

Verify the values ③

Error: smoke marked as “zsmoker” for
2 participants

zsmoker	<NA>
2	2

① Replace the errors (i.e.,
“zsmoker”) as missing values
using the **%in%** operator

② Replace “chinese” with
“Chinese” using the **%in%**
operator

③ Verify the changes

```
>
>
> # Ethnicity
> table(chp$ethnicity, useNA = 'ifany')
```

chinese	Chinese	Indians	Malays	<NA>
1	3882	3087	3049	3

“Chinese” spelled as
“chinese” for 1 participant

```
> chp$ethnicity[chp$ethnicity %in% "chinese"]
```

```
[1] "chinese"
```

```
> chp$ethnicity[chp$ethnicity %in% "chinese"] = "Chinese" ②
```

```
> table(chp$ethnicity, useNA = 'ifany')
```

Chinese	Indians	Malays	<NA>
3883	3087	3049	3

Verify the values ③

Data transformation

Transforming character variable to **binary variable**

Binary/dummy variable is one that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect.

Data transformation

Transforming character variable to **binary variable** (e.g. gender)

Binary/dummy variable is one that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect.

- ① Create new variable “female” using the **\$** operator
- ② Code “Female” as 1
- ③ Code “Male” as 0
- ④ Verify the values

Data transformation

Transforming character variable to **binary variable** (e.g. gender)

```
> table(chp$gender, useNA = 'ifany')
```

Female coded as 1

Female	Male	<NA>
5028	4990	4

Male coded as 0

① > chp\$female = NA

```
> table(chp$female, useNA = 'ifany')
```

<NA>

10022

② > chp\$female[chp\$gender %in% "Female"] = 1

③ > chp\$female[chp\$gender %in% "Male"] = 0

④ > table(chp\$female, useNA = 'ifany')

0	1	<NA>
4990	5028	4

Verify the values

- ① Create new variable “female” using the **\$** operator
- ② Code “Female” as 1
- ③ Code “Male” as 0
- ④ Verify the values

Data transformation

Alternative method to convert a character variable (e.g. smoke)

```
> # Smoke
> class(chp$smoke)          Converted from categorical to numeric
[1] "character"
> table(chp$smoke, useNA = 'ifany')      Smoking status has 4 unique
                                            values
```

	Daily smoker	Ex-smoker	Never-Smoker	Occasional smoker	<NA>
	1480	1120	6517	901	4

Data transformation

Alternative method to convert a character variable (e.g. smoke)

```
> # Smoke
> class(chp$smoke)
[1] "character"
```

Converted from categorical to numeric

```
> table(chp$smoke, useNA = 'ifany')
```

Smoking status has 4 unique values

Daily smoker	Ex-smoker	Never-Smoker	Occasional smoker	<code><NA></code>
1480	1120	6517	901	4

①

```
> chp$smoke1 = NA
> table(chp$smoke1, useNA = 'ifany')
```

`<NA>`

10022

②

```
> chp$smoke1[chp$smoke %in% "Daily smoker"] = 1
③ > chp$smoke1[chp$smoke %in% "Ex-smoker"] = 2
④ > chp$smoke1[chp$smoke %in% "Never-Smoker"] = 3
⑤ > chp$smoke1[chp$smoke %in% "Occasional smoker"] = 4
> table(chp$smoke1, useNA = 'ifany')
```

1	2	3	4	<code><NA></code>
1480	1120	6517	901	4

Verify the values

- ① Create new variable “smoke1” using the `$` operator
- ② Code “Daily smoker” as 1
- ③ Code “Ex-smoker” as 2
- ④ Code “Never-Smoker” as 3
- ⑤ Code “Occasional smoker” as 4
- ⑥ Verify the values

Data transformation : Re-categorising a categorical variable

Categorise variable `smoke` into binary variable [1: Daily smoker/Occasional smoker; and 0: Ex-smoker/Never-smoker]

```
> # Re-categorise variable smoke1 into binary variable
> chp$smoker = NA
> table(chp$smoker, useNA = 'ifany')
```

<NA>

10022

```
(1) > chp$smoker[chp$smoke %in% "Daily smoker"] = 1
(2) > chp$smoker[chp$smoke %in% "Occasional smoker"] = 1
(3) > chp$smoker[chp$smoke %in% "Ex-smoker"] = 0
(4) > chp$smoker[chp$smoke %in% "Never-Smoker"] = 0
(5) > table(chp$smoker, useNA = 'ifany')
```

0 1 <NA>

7637 2381 4

```
(6) > table(chp$smoke, chp$smoker, useNA = 'ifany')
```

	0	1	<NA>
Daily smoker	0	1480	0
Ex-smoker	1120	0	0
Never-Smoker	6517	0	0
Occasional smoker	0	901	0
<NA>	0	0	4

"Ex-smoker" and "Never-smoker" also correctly categorized into the non-smoker category

- ① Create new variable "smoker" using the `$` operator
- ② Code "Daily smoker" as 1
- ③ Code "Occasional smoker" as 1
- ④ Code "Ex-smoker" as 0
- ⑤ Code "Never-Smoker" as 0
- ⑥ Verify the values

Daily smoker and Occasional smoker correctly categorized into the smoker category

Data transformation : Categorising a continuous variable

Categorise variable bmi into binary variable [Overweight (Yes: $\geq 25 \text{ kg/m}^2$; No: $\leq 24.9 \text{ kg/m}^2$)]

```
> # Re-categorise variable bmi into binary variable [overweight (1:  $\geq 25 \text{ kg/m}^2$  ; 0:  $\leq 24.9 \text{ kg/m}^2$ )]  
> chp$overweight = NA  
> table(chp$overweight, useNA = 'ifany')  
  
<NA>  
10022
```

- ① Create new variable “overweight” using the **\$** operator
- ② Code “overweight” as 0 if BMI $< 25 \text{ kg/m}^2$
- ③ Code “overweight” as 1 if BMI $\geq 25 \text{ kg/m}^2$
- ④ Verify the range

Data transformation : Categorising a continuous variable

Categorise variable bmi into binary variable [Overweight (Yes: $\geq 25 \text{ kg/m}^2$; No: $\leq 24.9 \text{ kg/m}^2$)]

```

① > # Re-categorise variable bmi into binary variable [overweight (1:  $\geq 25 \text{ kg/m}^2$  ; 0:  $\leq 24.9 \text{ kg/m}^2$ )]
> chp$overweight = NA
> table(chp$overweight, useNA = 'ifany')

<NA>
10022

② > chp$overweight[chp$bmi_drop < 25] = 0
③ > chp$overweight[chp$bmi_drop >= 25] = 1
> #verify the range
> summary(chp$bmi_drop)
   Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
   15.10    22.10   25.20  25.37   28.50  43.50     3

④ > summary(chp$bmi_drop[chp$overweight %in% 0])
   Min. 1st Qu. Median Mean 3rd Qu. Max.
   15.10    19.80   22.00  21.51   23.60  24.90

> summary(chp$bmi_drop[chp$overweight %in% 1])
   Min. 1st Qu. Median Mean 3rd Qu. Max.
   25.00    26.50   28.40  28.95   30.80  43.50

```

- ① Create new variable “overweight” using the **\$** operator
- ② Code “overweight” as 0 if BMI $< 25 \text{ kg/m}^2$
- ③ Code “overweight” as 1 if BMI $\geq 25 \text{ kg/m}^2$
- ④ Verify the range

Check the range for each group

Your task (~2 mins)

- How many participants are in the overweight category?

Drop observations with erroneous and/or missing data

Before deleting the observations with missing and/or erroneous data

The original (raw) file should remain *untouched* as an .xls document – Never overwrite!

Store this file as “chp_processed_data”

```
chp_processed_data = chp
```

Drop observations with erroneous and/or missing data

Before deleting the observations with missing and/or erroneous data

The original (raw) file should remain *untouched* as an .xls document – Never overwrite!

Store this file as “chp_processed_data”

```
chp_processed_data = chp
```

- ① Identify rows with missing and/or erroneous data using the “which” function
- ② Delete the rows with missing and/or erroneous data
- ③ Save cleaned data

```
(1) > # Identify the rows with missing and/or erroneous data using the "which" function
> problemrows = which(is.na(chp$age_drop) | is.na(chp$gender) | is.na(chp$bmi_drop) | is.na(chp$ethnicity) | is.na(chp$smoke) | is.na(chp$cvd) | is.na(chp$ldl_drop))
> chp[problemrows,]
# A tibble: 22 x 15
   id age gender bmi ethnicity smoke      cvd    ldl ldl_drop age_drop bmi_drop female smoke1 smoker overweight
   <dbl> <dbl> <chr> <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
 1 10001    4 Female  24.9 Malays  Never-Smoker  0  3.09   3.09     NA   24.9    1     3     0     0
 2 10002   141 Male   22.8 Indians Never-Smoker  0  2.79   2.79     NA   22.8    0     3     0     0
 3 10003   150 Female 19.6 Indians Never-Smoker  0  2.76   2.76     NA   19.6    1     3     0     0
 4 10004     2 Female  21.5 Chinese Never-Smoker  0  3.55   3.55     NA   21.5    1     3     0     0
 5 10005    82 Female  22.8 Malays  NA          1  2.3     2.3    82   22.8    1     NA    NA     0
 6 10006    59 Female  NA        Chinese Never-Smoker  0  4       4     59   NA      1     3     0     NA
 7 10007    65 Male   62.1 Malays  Never-Smoker  1  2.32   2.32    65   NA      0     3     0     NA
 8 10008    64 Male   59.2 Malays  NA          1  2.87   2.87    64   NA      0     NA    NA     NA
 9 10009    70 Male   26.7 Indians Never-Smoker  0  -2.34   NA     70   26.7    0     3     0     1
10 10010   NA        Male   29.2 Chinese Never-Smoker  0  11.8    NA     NA   29.2    0     3     0     1
# ... with 12 more rows
> # Delete the rows with missing and/or erroneous data
> chp = chp[-problemrows,]
> # delete this variable as it is no longer needed
> rm(problemrows)
```

Drop observations with erroneous and/or missing data

Before deleting the observations with missing and/or erroneous data

The original (raw) file should remain *untouched* as an .xls document – Never overwrite!

Store this file as “chp_processed_data”

```
chp_processed_data = chp
```

- ① Identify rows with missing and/or erroneous data using the “which” function
- ② Delete the rows with missing and/or erroneous data
- ③ Save cleaned data

```
> # Save cleaned data: as Excel csv file (easy to open this file in MS Excel)  
③ > write.csv(chp,file ='data/chp_cleaned.csv',row.names = FALSE)  
>  
> # Save cleaned data: as R data file (easy to load this file into R for further analysis)  
> save(chp,file = 'data/chp_cleaned.rdata')  
>  
> # Save cleaned data: as RDS file (easy to load this file into R for further analysis)  
> saveRDS(chp,file = 'data/chp_cleaned.rds')
```



Key takeaway messages

- R is a very flexible statistical software
- There are many ways to achieve the same output, but not all ways are equally efficient (be careful)
- The online resources such as Stack Overflow, ...

Thank you