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! ☺

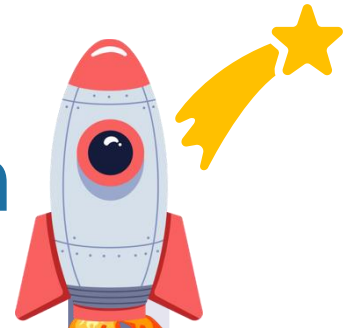
Biostatistics for Public Health

Share a little about yourself

TM-CM02

Biostatistics for Public Health

- quizzes
- 2 assignments



4

Logistic regression

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

3

Linear regression

Data prep, descriptive & inferential statistics, linear regression

2

Hypothesis testing

Inferential statistics

1

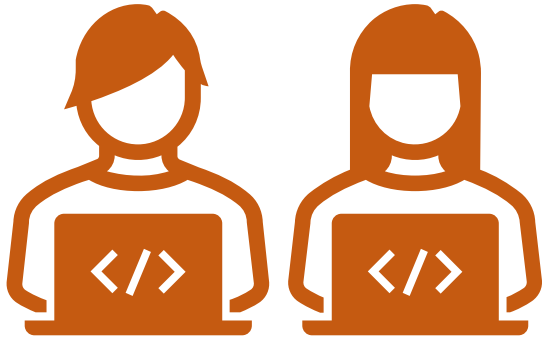
Categorical data analysis

Data prep, descriptive & inferential statistics

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

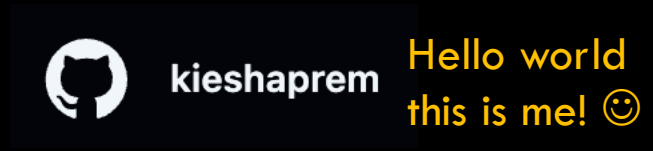


15-min Snack/bio break

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



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).

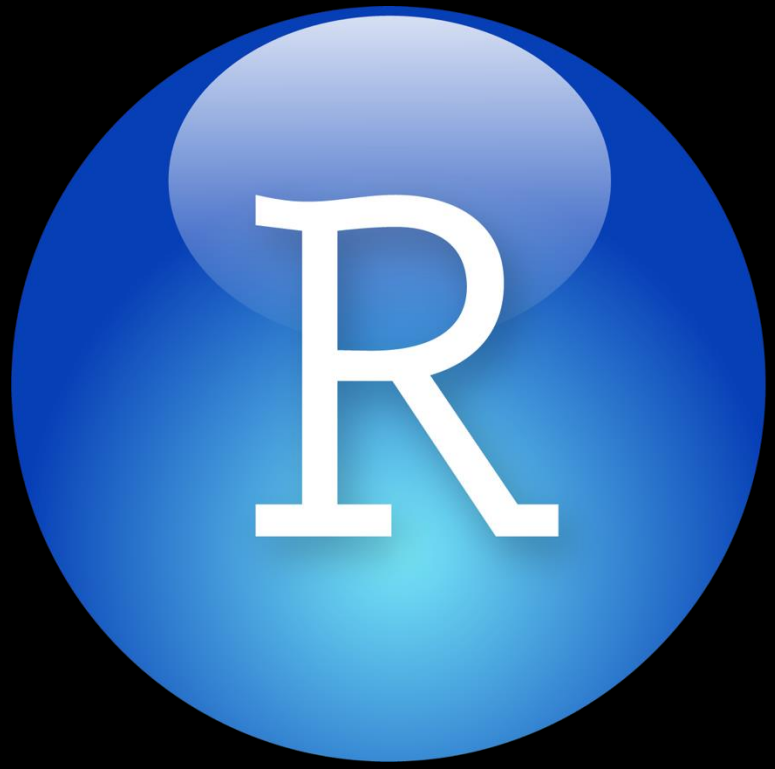
1. **Download & install R: https://cran.r-project.org/**
2. **Download & install RStudio: https://posit.co/download/rstudio-desktop/**

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/

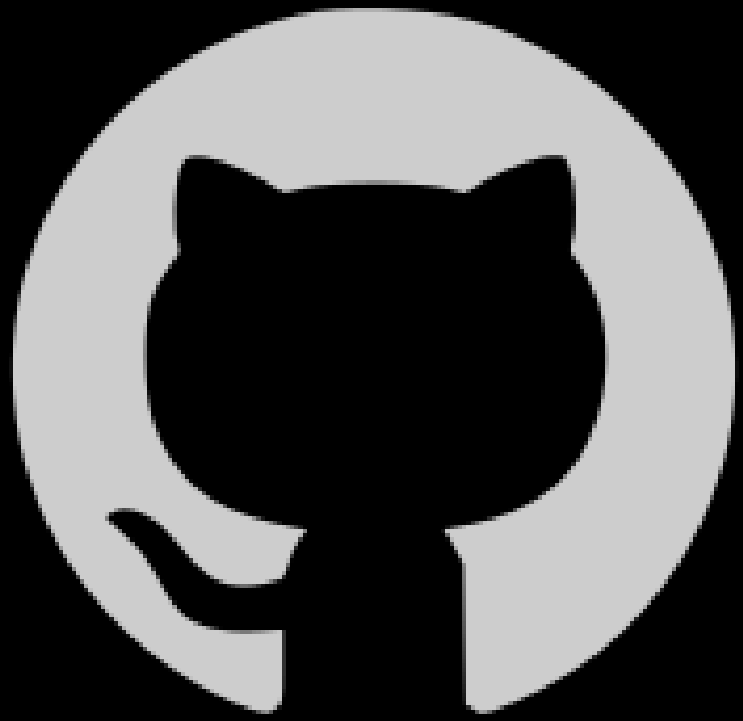
2. **Download GitHub Desktop:** Visit https://desktop.github.com/download/ and download the appropriate version for your operating system (Windows or macOS).

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

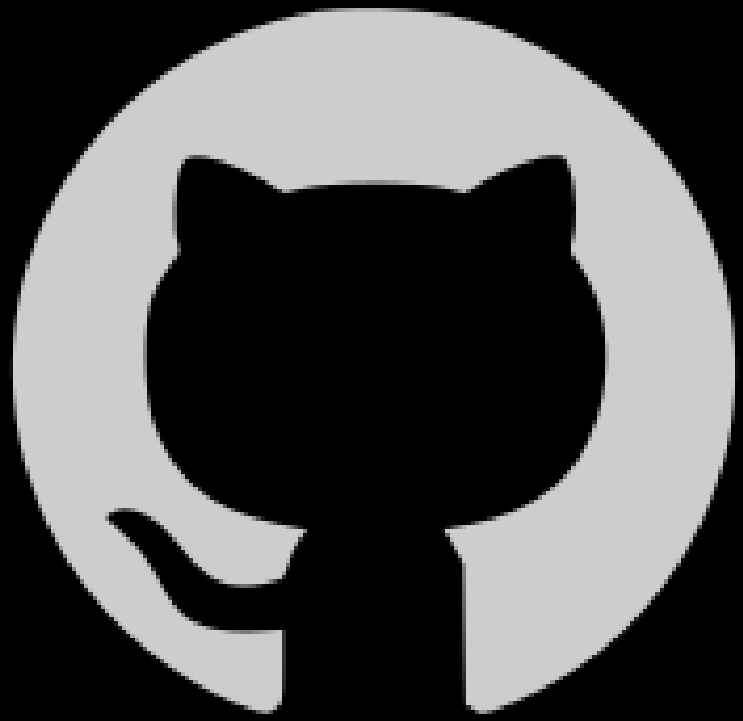
4. **Sign in to GitHub:** Once installed, open the GitHub Desktop application. You will be prompted to sign in to your GitHub account (step 1).

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

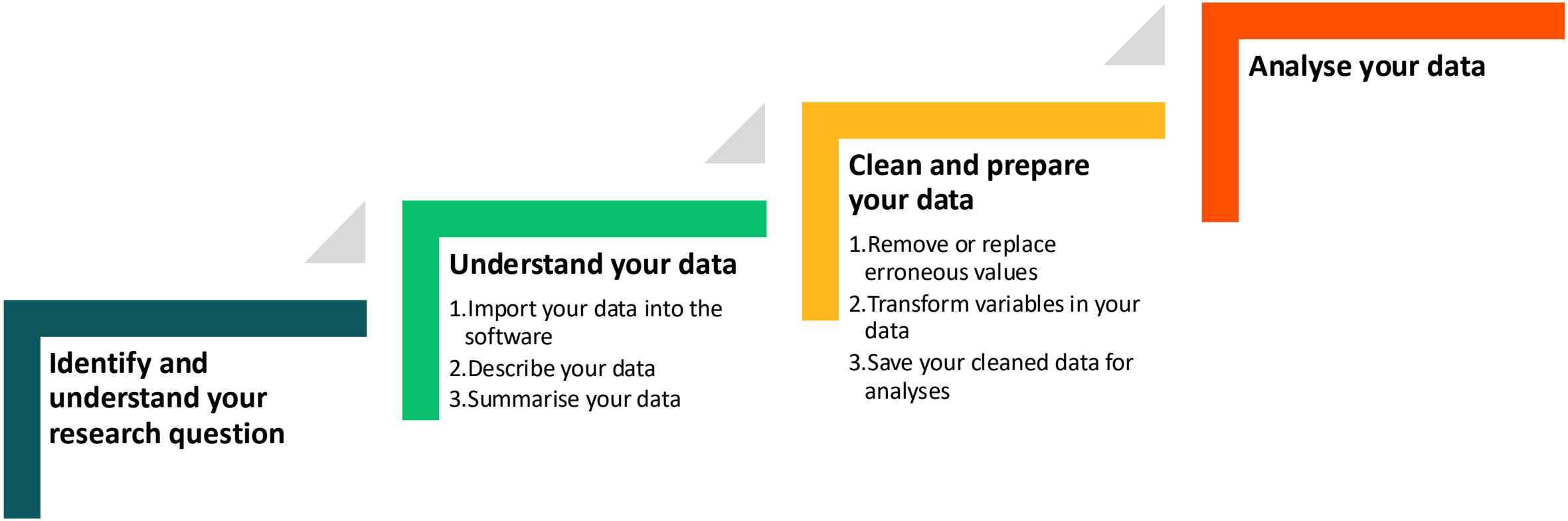
TM-CM02 repository

GNU General Public License v3.0 Updated 4 hours ago

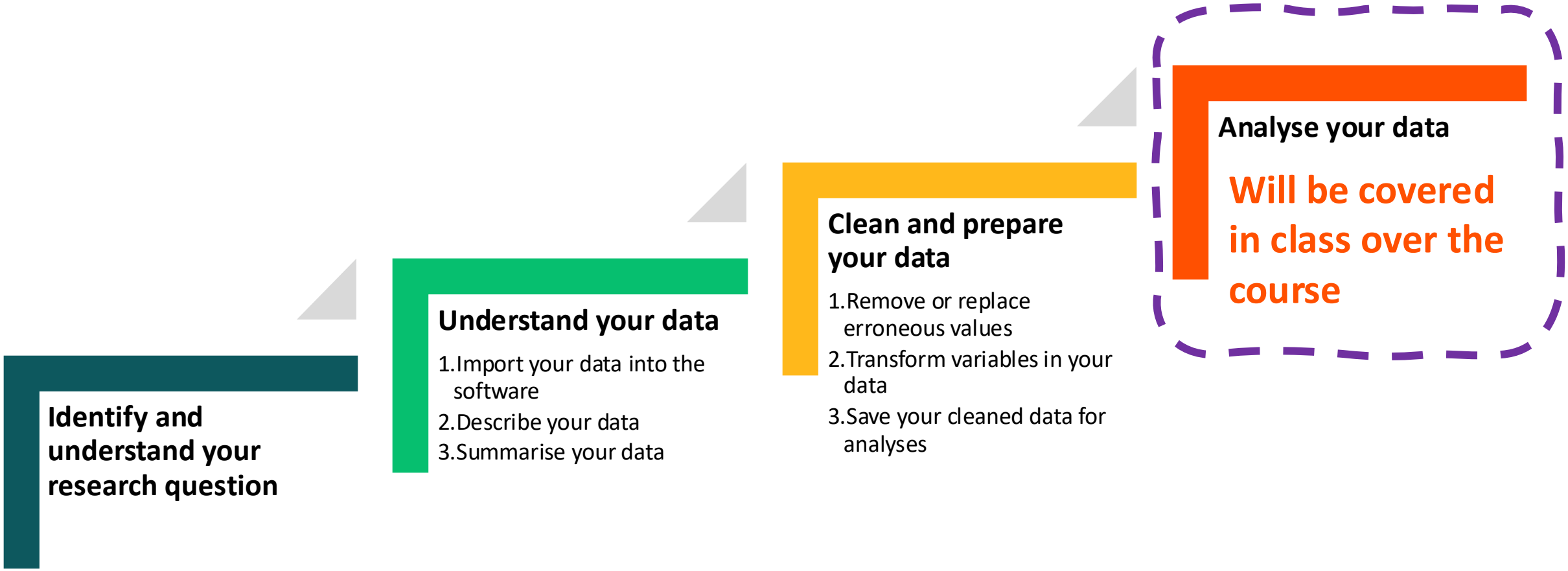
15-min Snack/bio break

https://youtu.be/iGjv-igK4ll

Data analysis



Data analysis



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 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

	A	B	C	D	E	F	G	H
1	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
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	3.40
10024								

Total number of participants is 10,022.

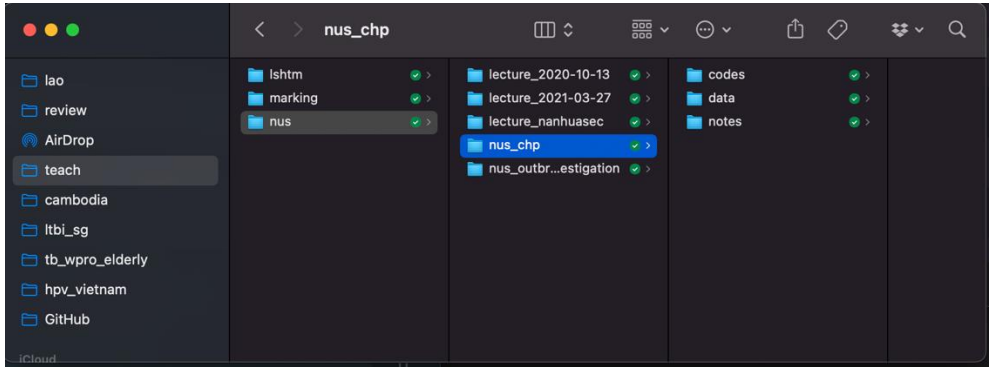
R interface

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

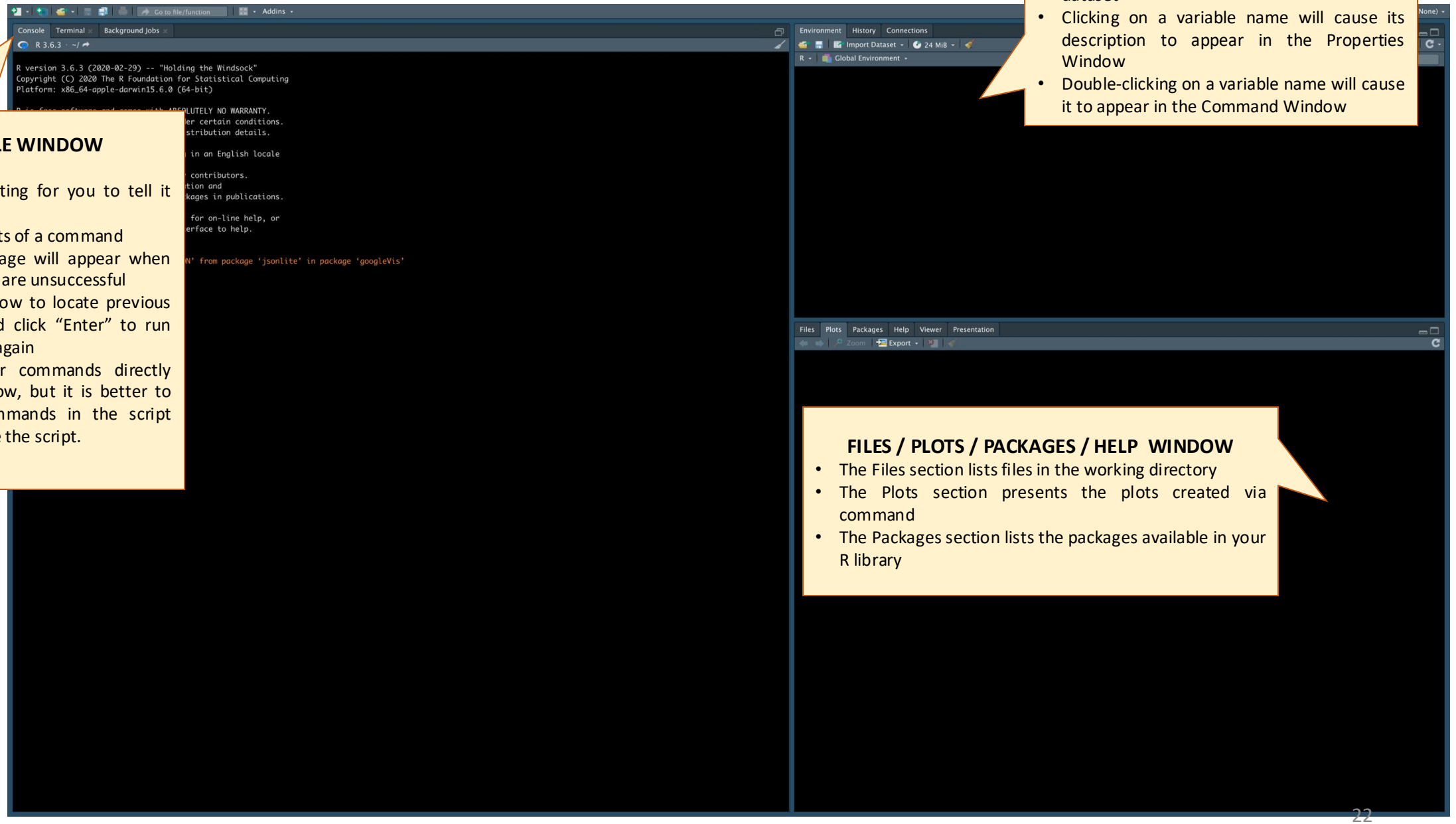
1. the R software itself, and
2. RStudio Desktop (we will only need to open this software, but RStudio needs R installed to work)

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



CONSOLE WINDOW

- Here, R is waiting for you to tell it what to do,
- Show the results of a command
- An error message will appear when the commands are unsuccessful
- use the up arrow to locate previous commands and click “Enter” to run the command again
- You can enter commands directly into this window, but it is better to enter the commands in the script editor and save the script.

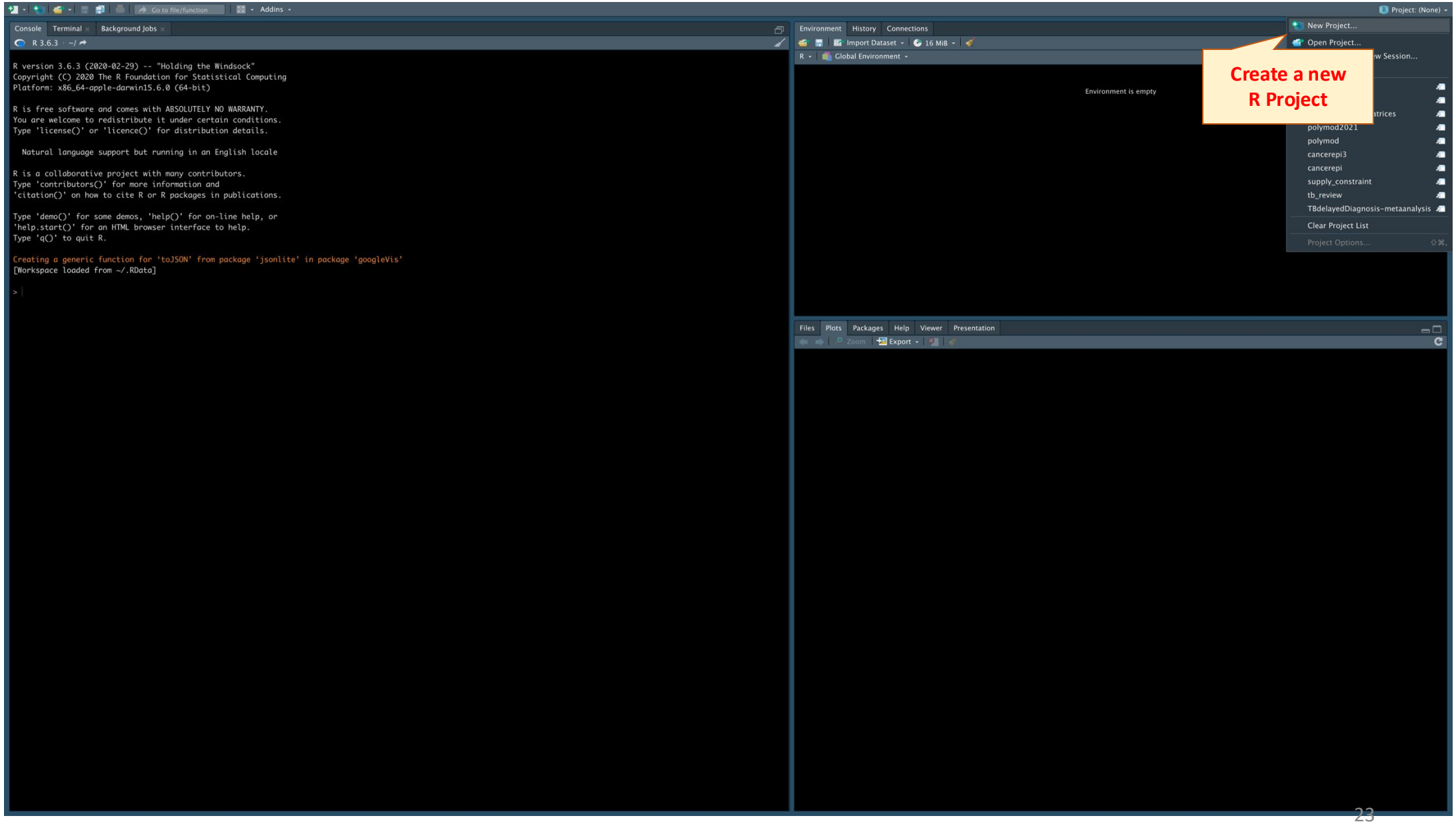
ENVIRONMENT WINDOW

- Variables in the dataset will be listed with their labels in the order they appear on the dataset
- Clicking on a variable name will cause its description to appear in the Properties Window
- Double-clicking on a variable name will cause it to appear in the Command Window

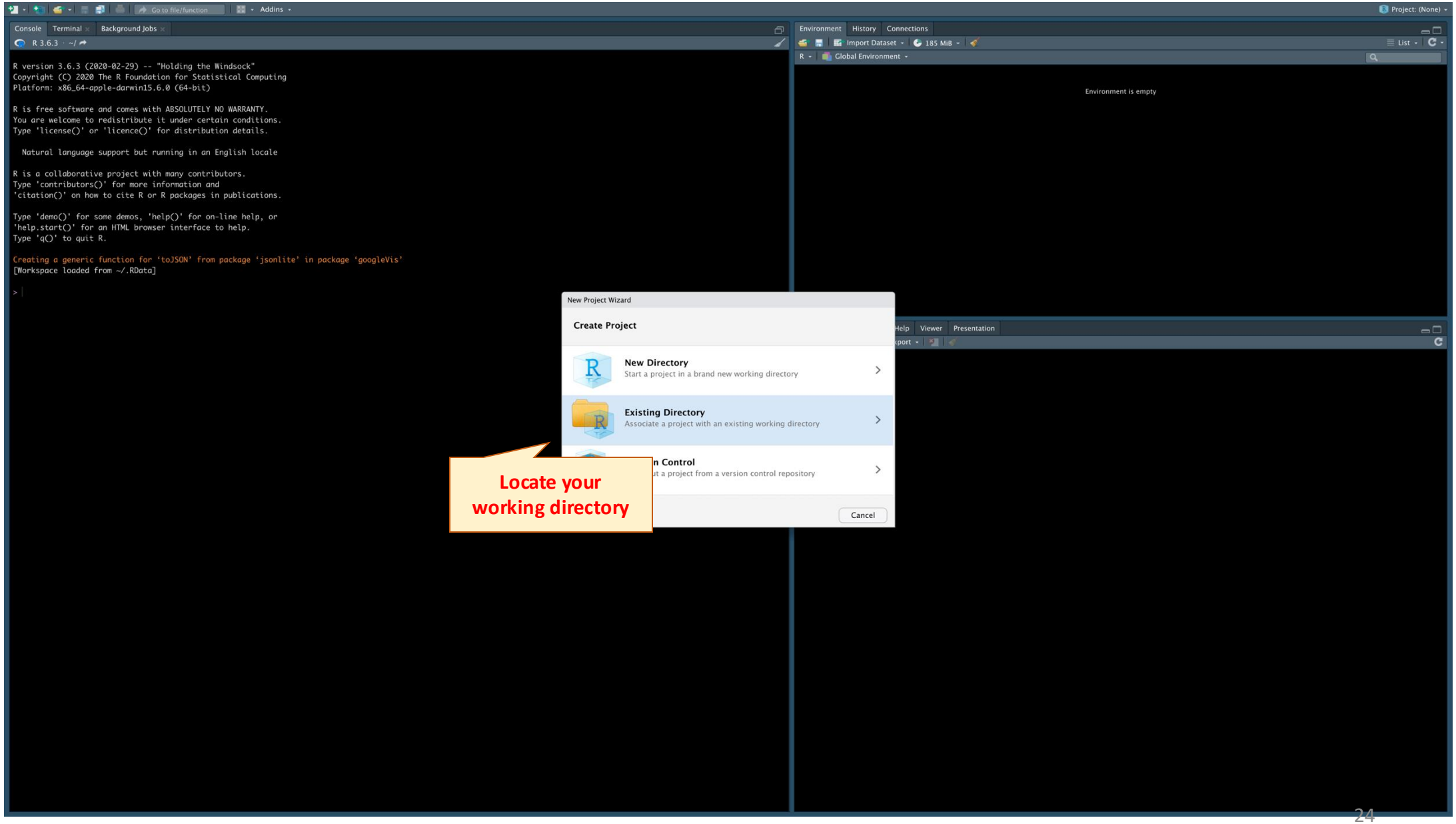
FILES / PLOTS / PACKAGES / HELP WINDOW

- The Files section lists files in the working directory
- The Plots section presents the plots created via command
- The Packages section lists the packages available in your R library

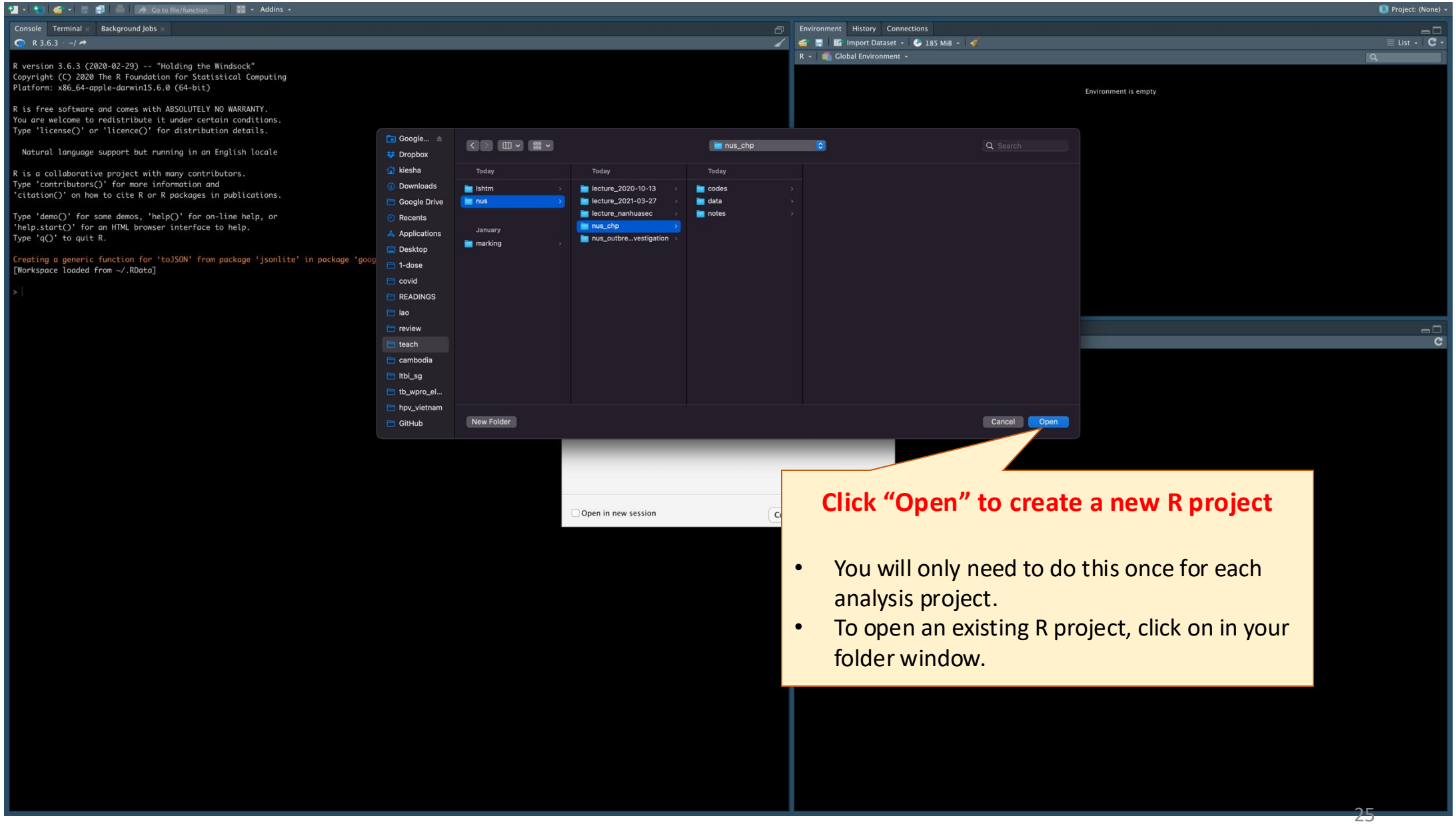
R interface – R Project (optional, but strongly recommended)



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



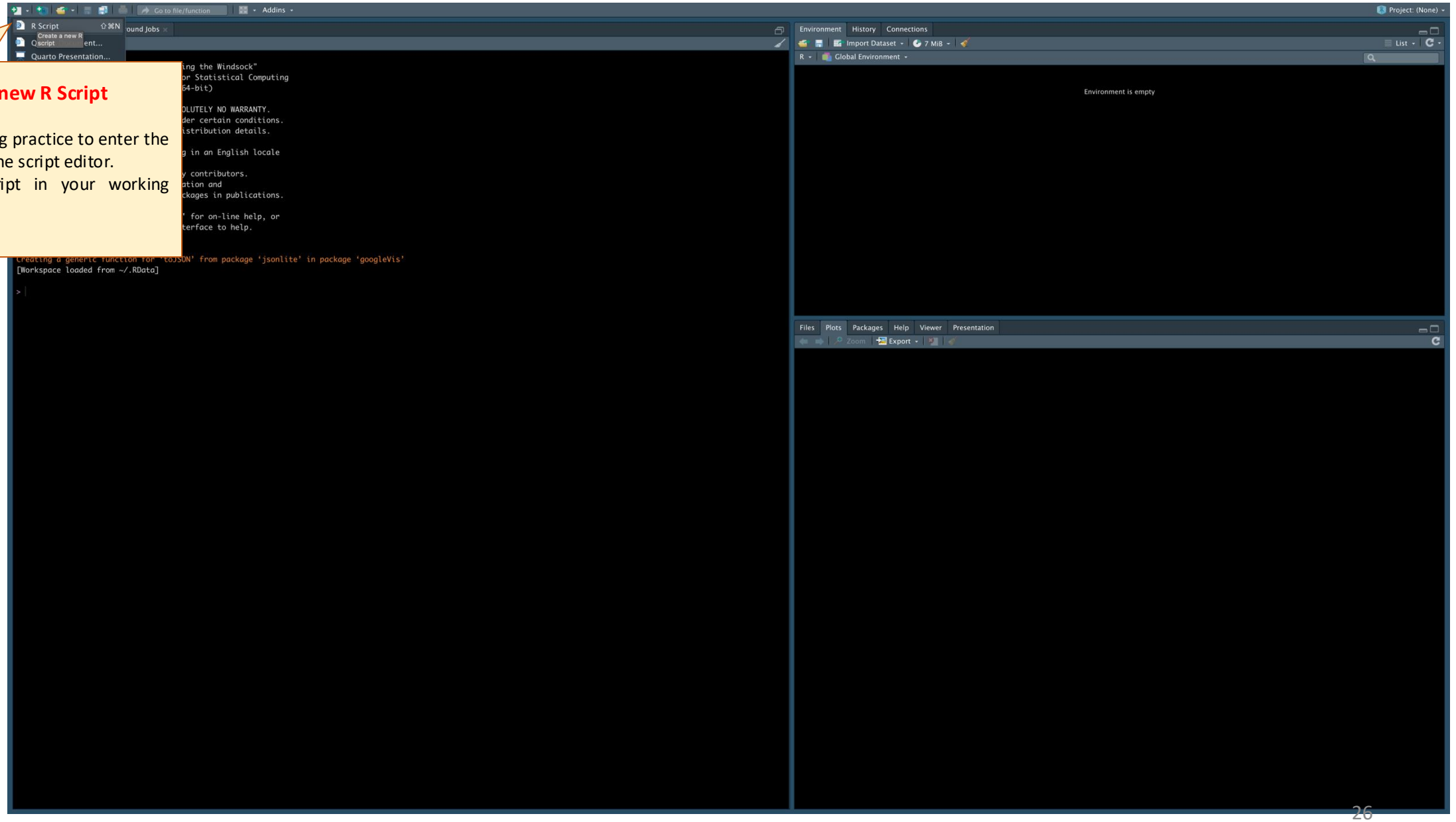
R interface – R Project (optional, but strongly recommended)



R interface – use scripts to write and save codes

Create a new R Script

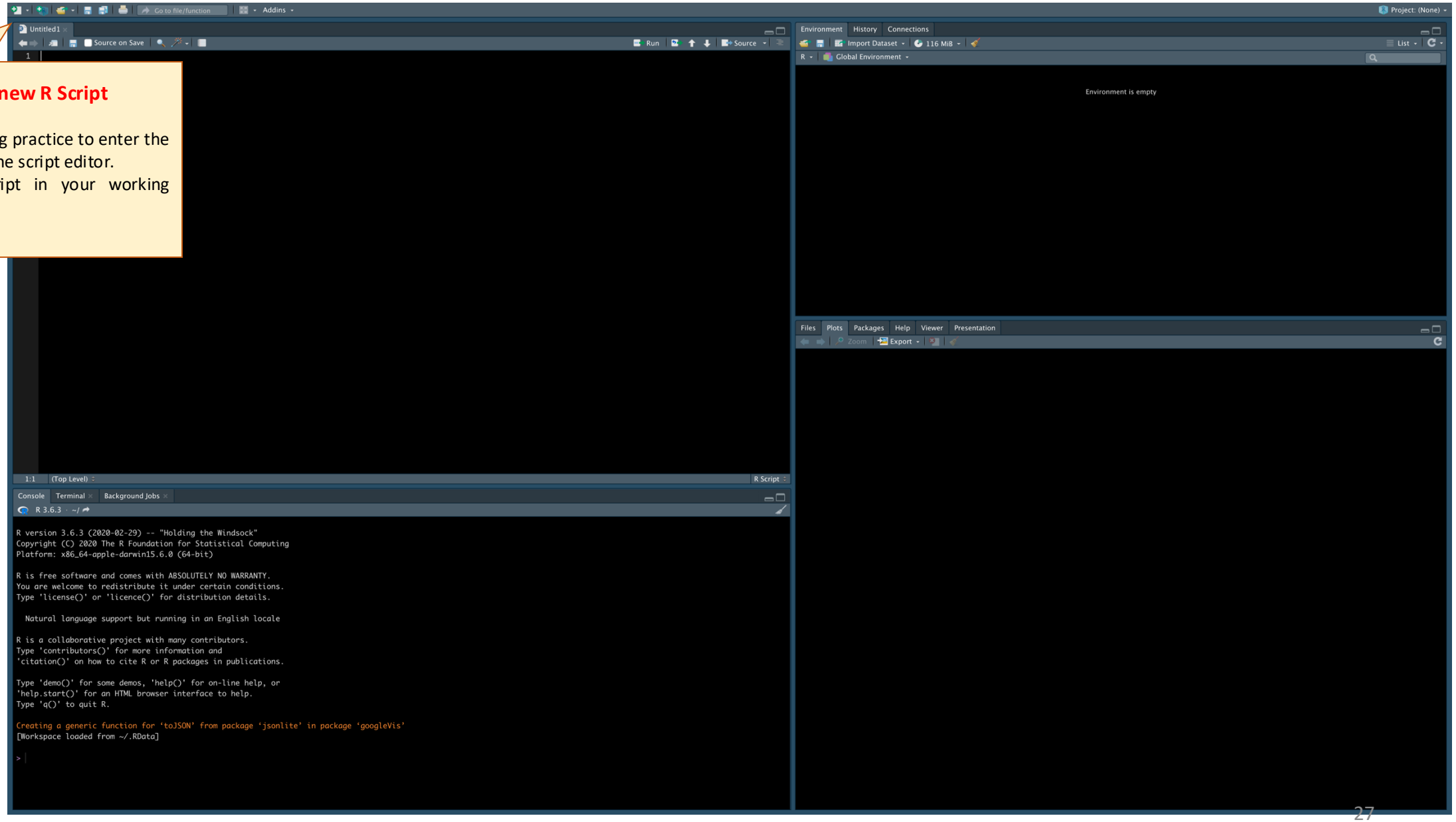
- It is good coding practice to enter the commands in the script editor.
- Save your script in your working directory.



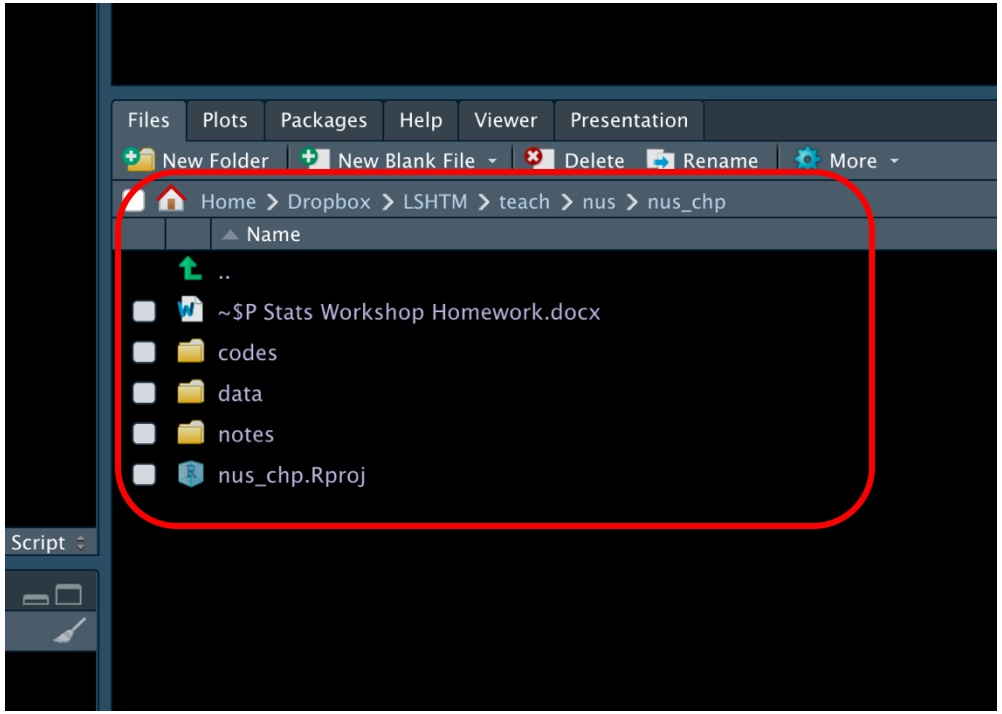
R interface – use scripts to write and save codes

Create a new R Script

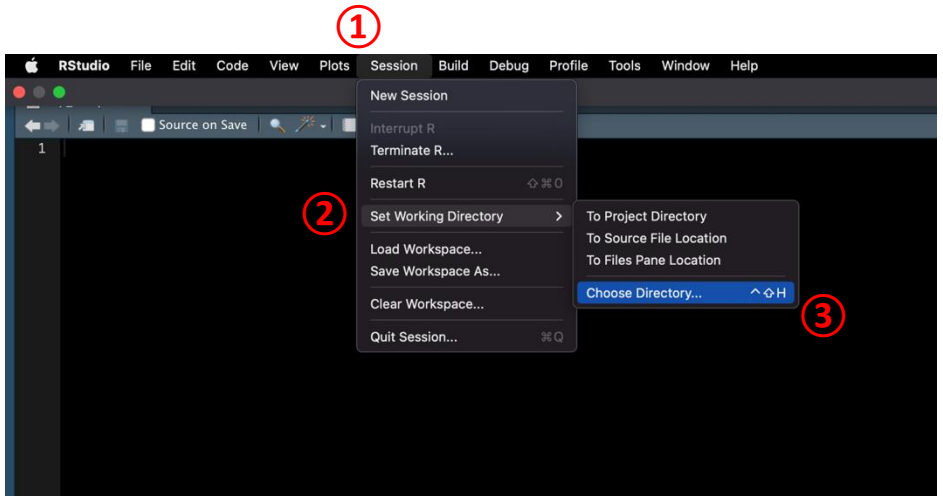
- It is good coding practice to enter the commands in the script editor.
- Save your script in your working directory.



R 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:

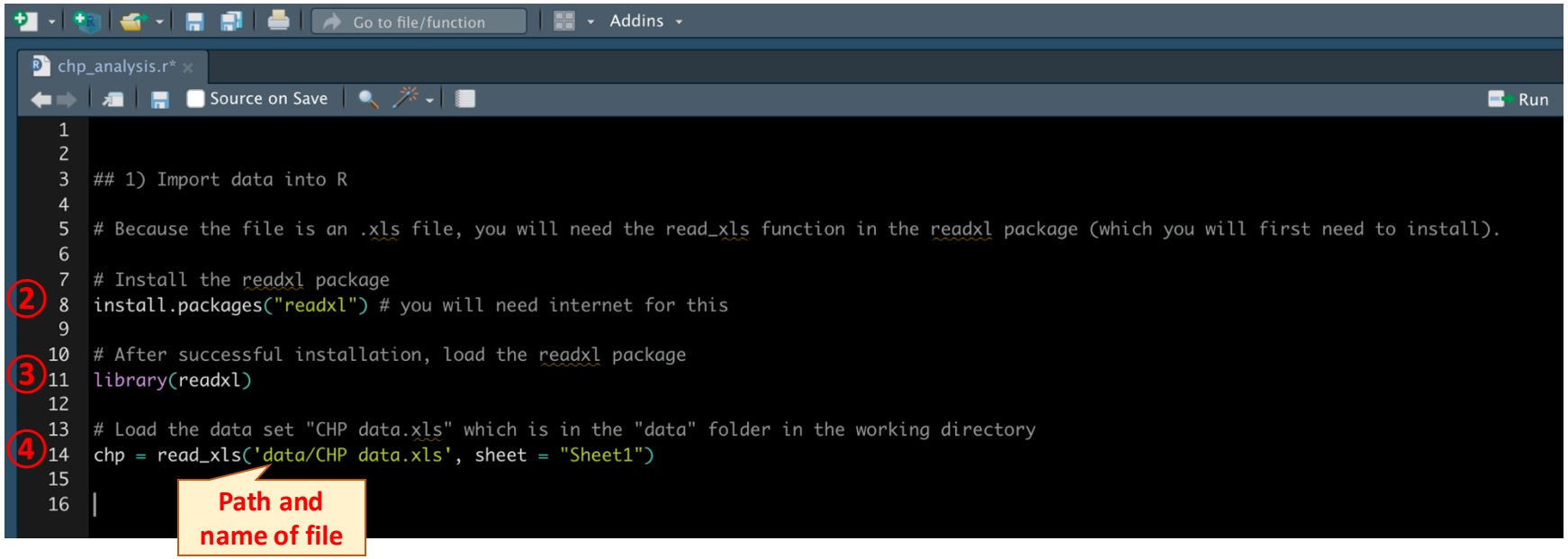
```
getwd()
```

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



Commenting

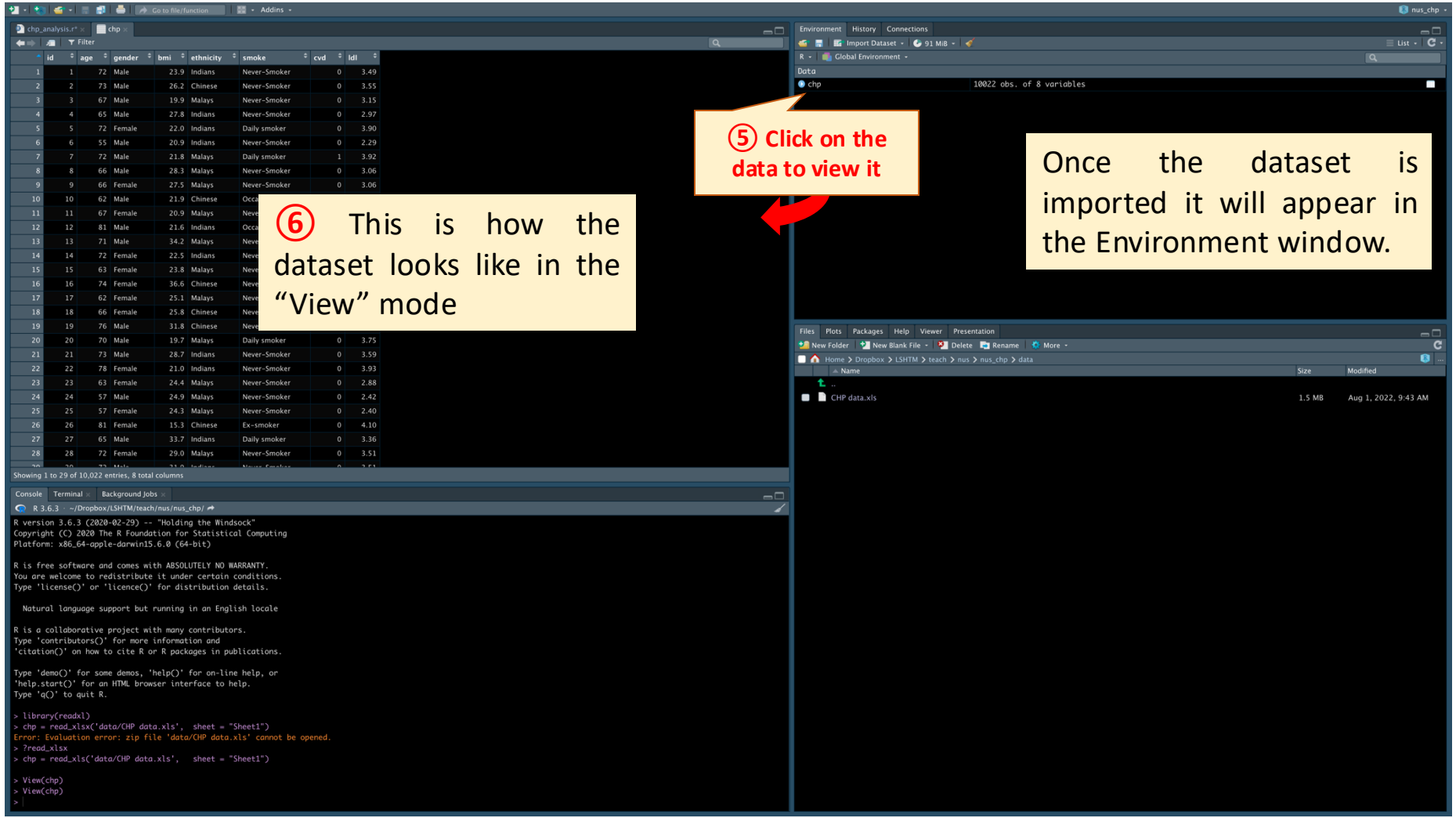
- Use # signs to comment.
- *Comment liberally* in your R scripts.
- Anything to the right of a # is ignored by R.

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



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

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

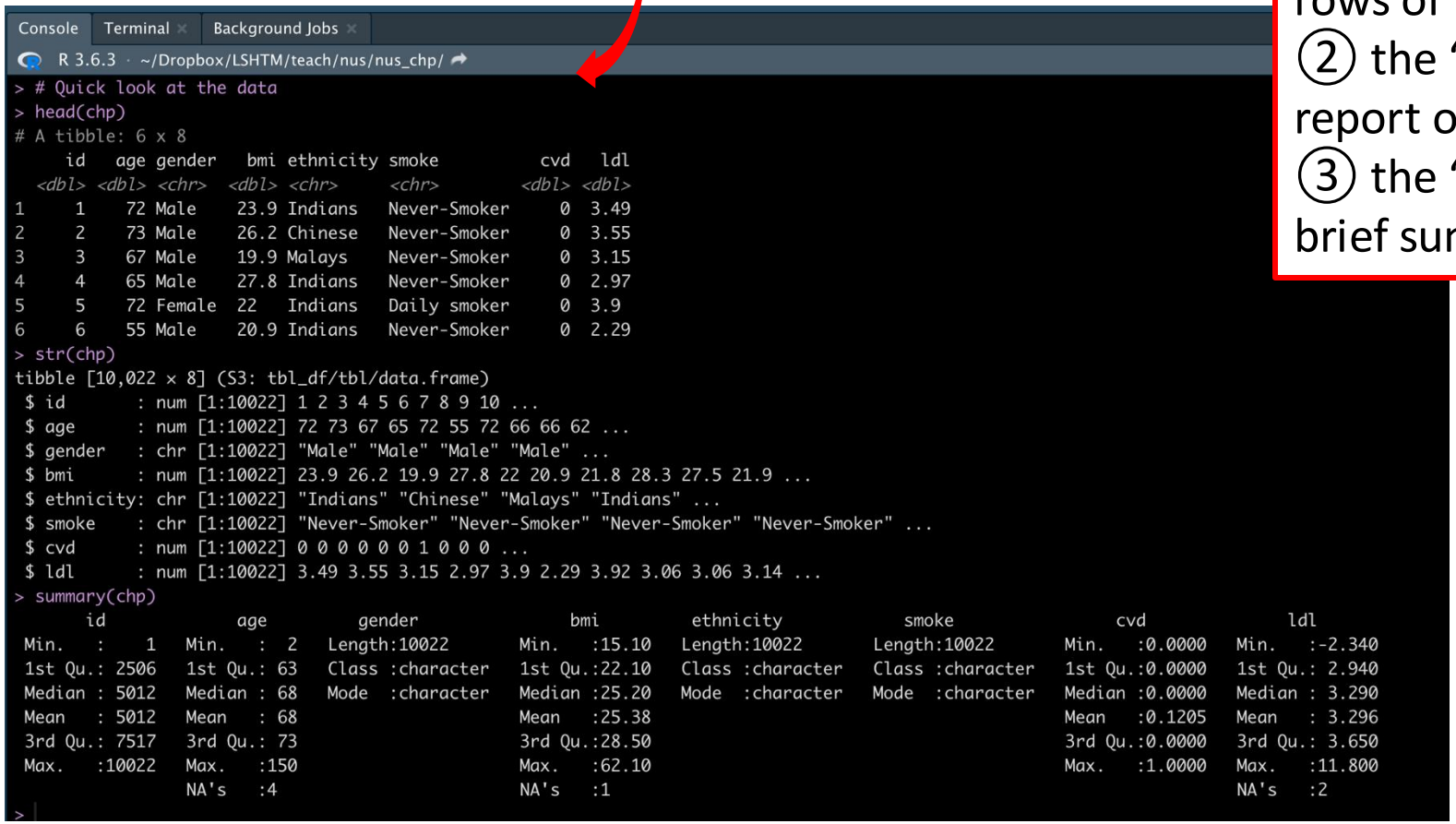
STOP
Something is not right!

UNCLEAR
Let's review later!

OK
Good to go!

Describing the data

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



```
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 × 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                      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
>
```

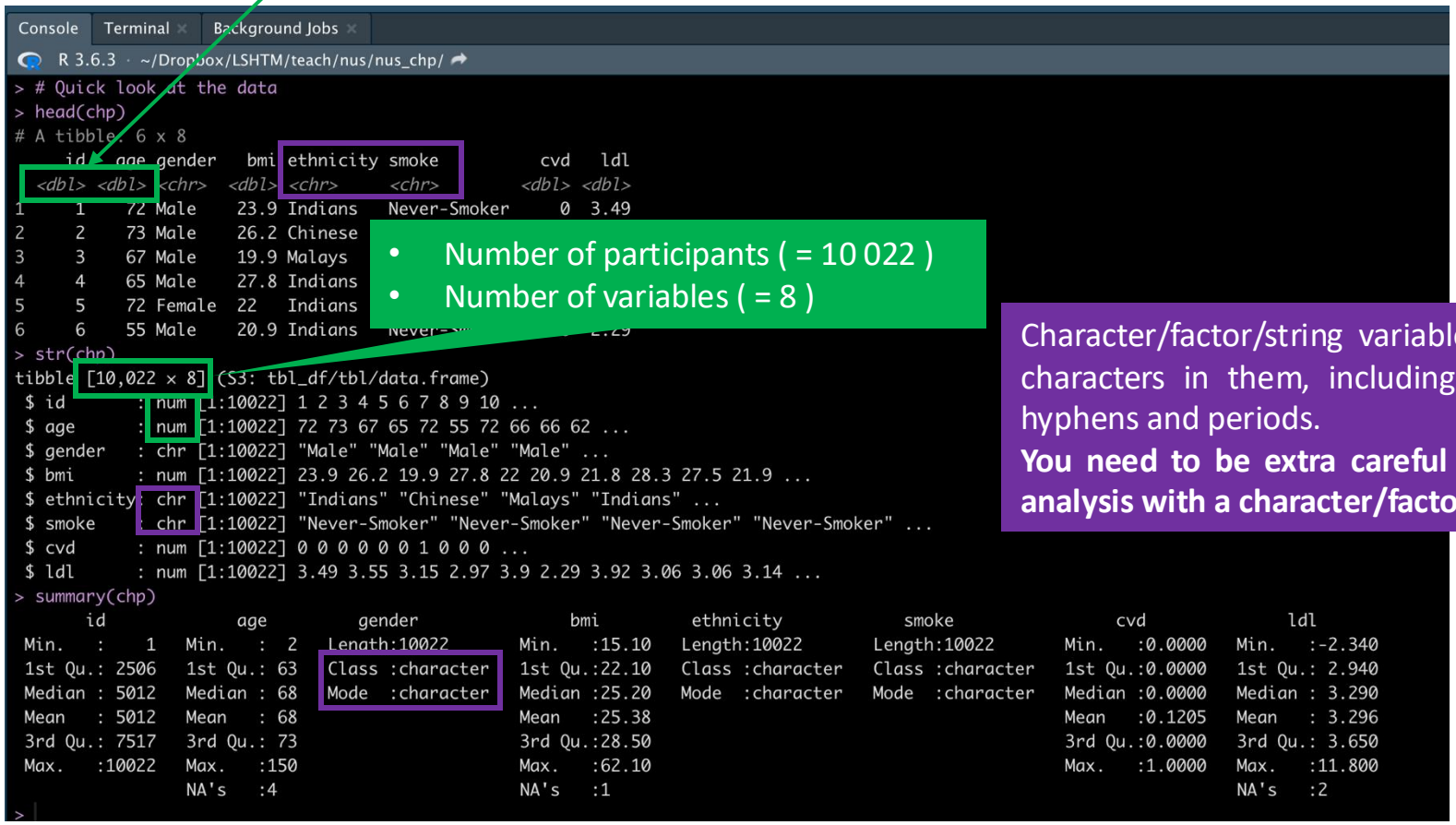
① the “head” function prints the top 6 rows of the data
② the “str” function provides a compact report on all variables in the dataset
③ the “summary” function provides a brief summary of the variables

Your task (~2 mins)

- Review the output

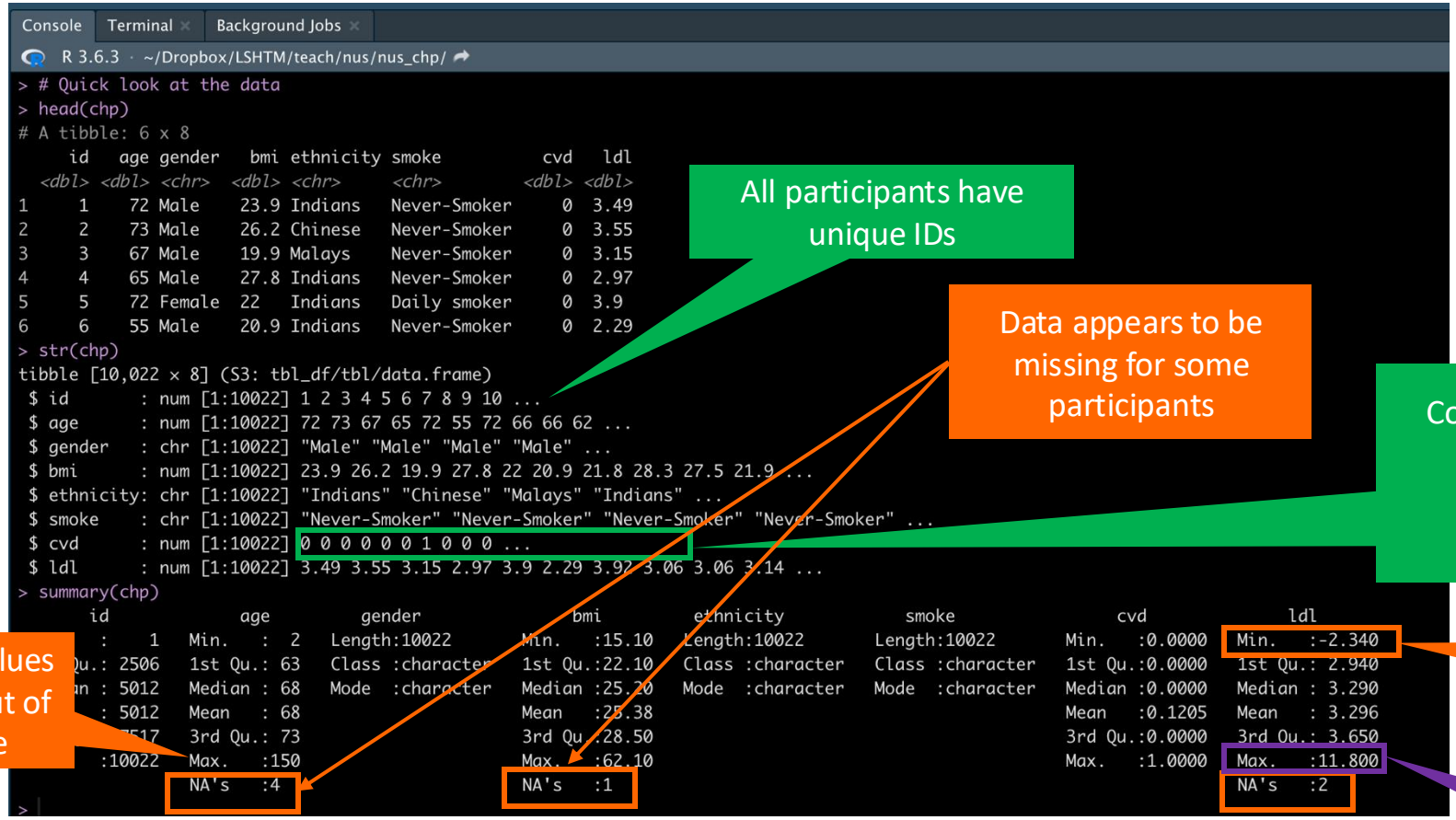
Describing the data: review the output

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



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



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 × 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                      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
>
```

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

chinese Chinese Indians  Malays    <NA>
      1    3882    3087    3049       3
> table(chp$smoke, useNA = 'ifany')

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

Gender should have 2 unique values: Male and Female

Data appears to be missing for some participants

Ethnicity should have 3 unique values

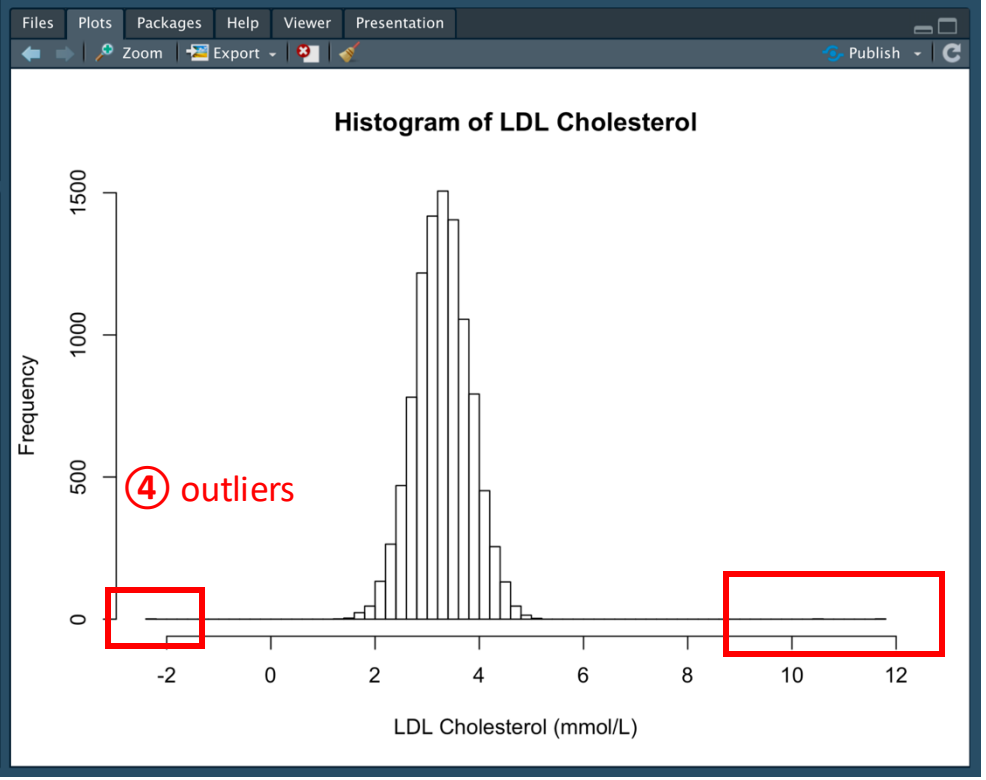
Smoke should have 4 unique values

Summarising the data (continuous variables)

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)

```
chp_analysis.r*
Source on Save | Run | Source

summary(chp)

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


# Summarize the data: continuous variable
summary(chp$ldl)
quantile(chp$ldl)
quantile(chp$ldl,na.rm = TRUE)
hist(chp$ldl,breaks = 100,xlab="LDL Cholesterol (mmol/L)", main="Histogram of LDL Cholesterol")




summary(chp$bmi)
table(chp$ethnicity, useNA = 'ifany')
table(chp$smoke, useNA = 'ifany')


prop.table(table(chp$gender))

37:1 (Top Level) R Script
```

```
Console | Terminal | Background Jobs
R 3.6.3 · ~/Dropbox/LSHTM/teach/nus/nus_chp/
> # Summarize the data: continuous variable
> summary(chp$ldl)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 -2.340   2.940   3.290   3.296   3.650  11.800       2   ①
> quantile(chp$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
-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")   ③
>
```



Summarising the data (continuous variables)

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)

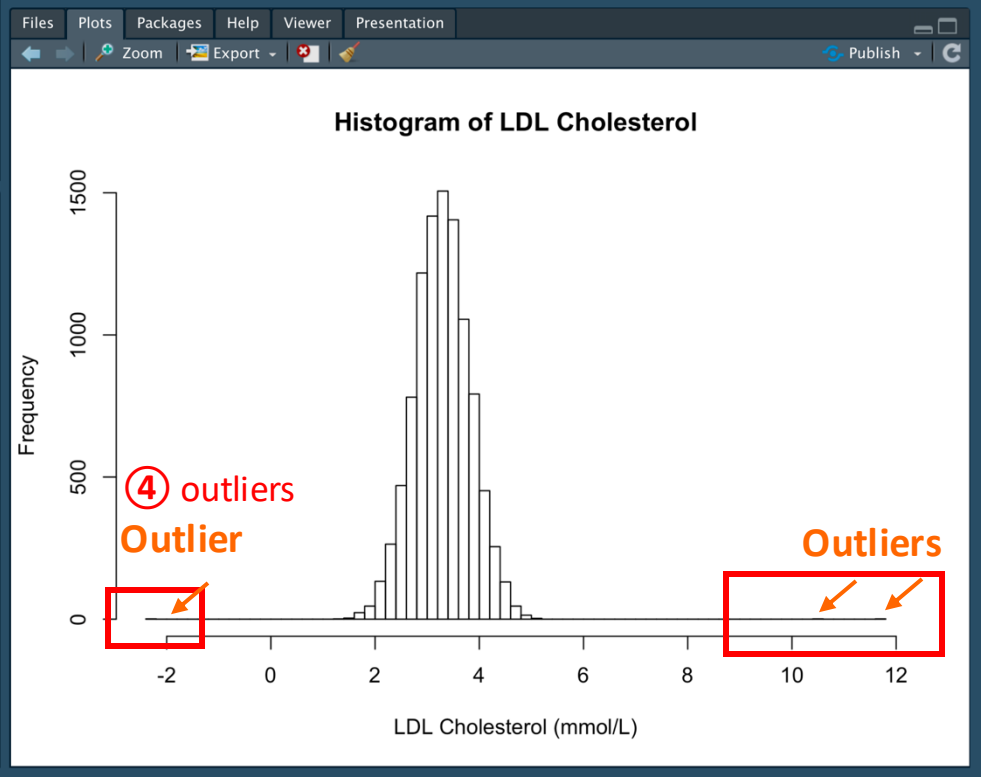
```
chp_analysis.r*
Source on Save | Run | Source
summary(chp)

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


# Summarize the data: continuous variable
summary(chp$ldl)
quantile(chp$ldl)
quantile(chp$ldl,na.rm = TRUE)
hist(chp$ldl,breaks = 100,xlab="LDL Cholesterol (mmol/L)", main="Histogram of LDL Cholesterol")




summary(chp$bmi)
table(chp$ethnicity, useNA = 'ifany')
table(chp$smoke, useNA = 'ifany')


prop.table(table(chp$gender))
37:1                                                    R Script
```

Some values appear to be too high

Erroneous data : negative values

```
R 3.6.3 · ~/D                    acn/nus/nus_chp/
> # Summari          data: continuous variable
> summary(chp$ldl)
    Min.  1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  -2.340   2.940   3.290   3.296   3.650  11.800       2   ①
> quantile(chp$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
-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")  ③
>
```

LDL values missing for 2 participants



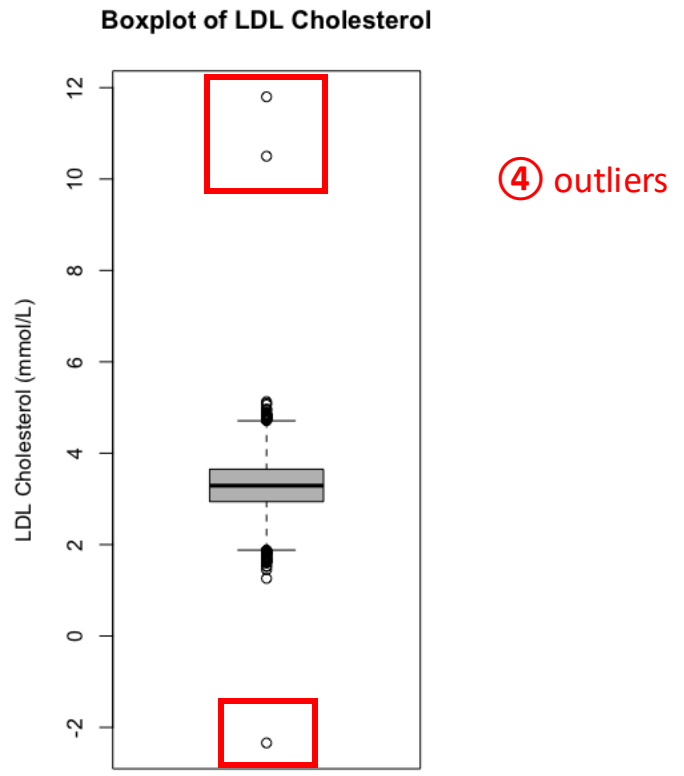
Summarising the data (continuous variables)

```
summary(chp)

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


# Summarize the data: continuous variable
summary(chp$ldl)
quantile(chp$ldl)
quantile(chp$ldl,na.rm = TRUE)
hist(chp$ldl,breaks = 100,xlab="LDL Cholesterol (mmol/L)", main="Histogram of LDL Cholesterol")









summary(chp$bmi)
table(chp$ethnicity, useNA = 'ifany')
table(chp$smoke, useNA = 'ifany')


prop.table(table(chp$gender))
```

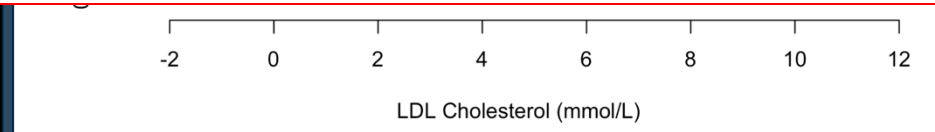
```
> # Summarize the data: continuous variable
> summary(chp$ldl)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 -2.340   2.940   3.290   3.296   3.650  11.800       2
> quantile(chp$ldl)
Error in quantile.default(chp$ldl) :
  missing values and NaN's not allowed if 'na.rm' is FALSE
> quantile(chp$ldl,na.rm = 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")
>
```

Plot the **box plot**

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



④ outliers



In-class Exercise #1

Summarize the data for the following continuous variables:

(i) Age

(ii) BMI

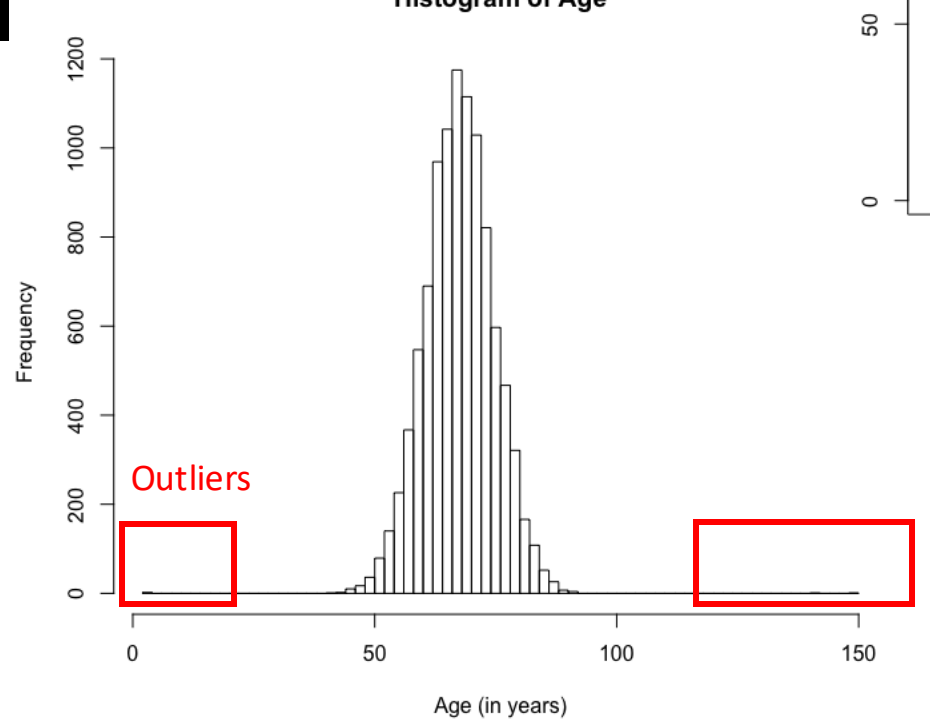
Do you observe any missing/erroneous values?

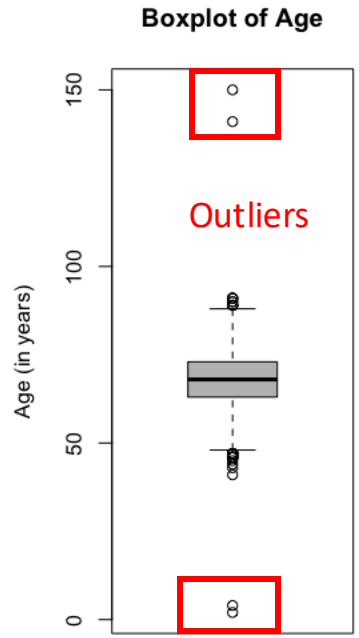
In-class Exercise #1 Solution

```
> # Age
> summary(chp$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      2      63      68      68      73     150       4
> quantile(chp$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

Age missing for some participants



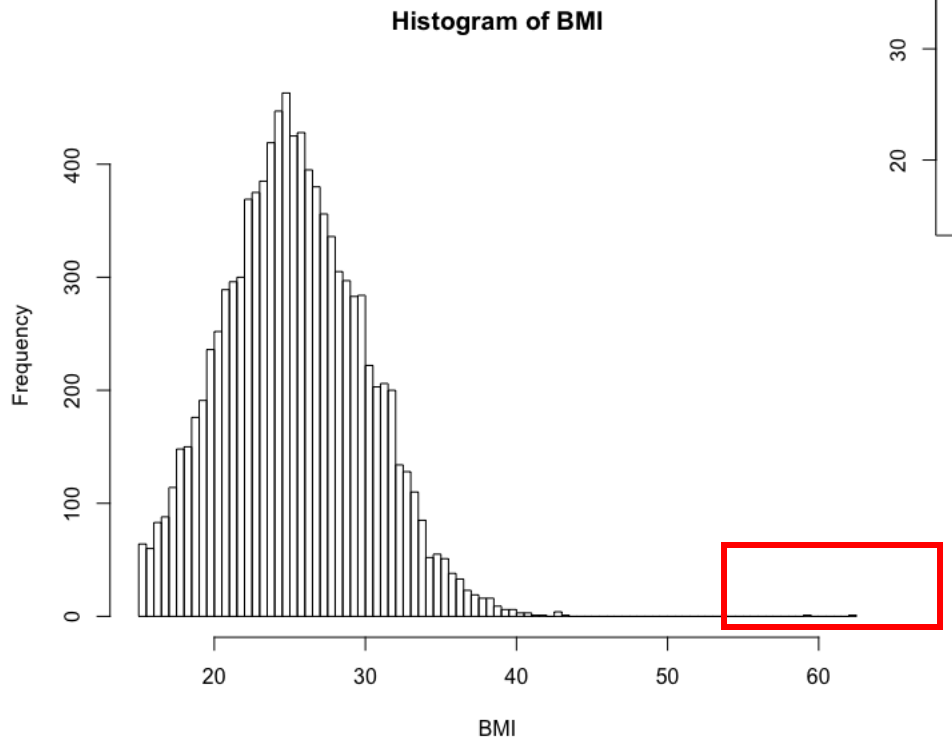


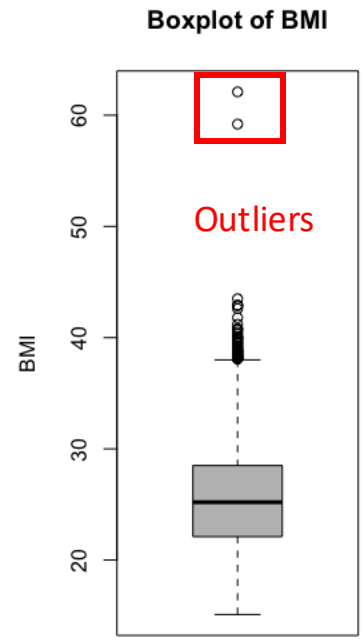
In-class Exercise #1 Solution

Some values appear to be too high

BMI missing for some participants

```
> # 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')
>
```





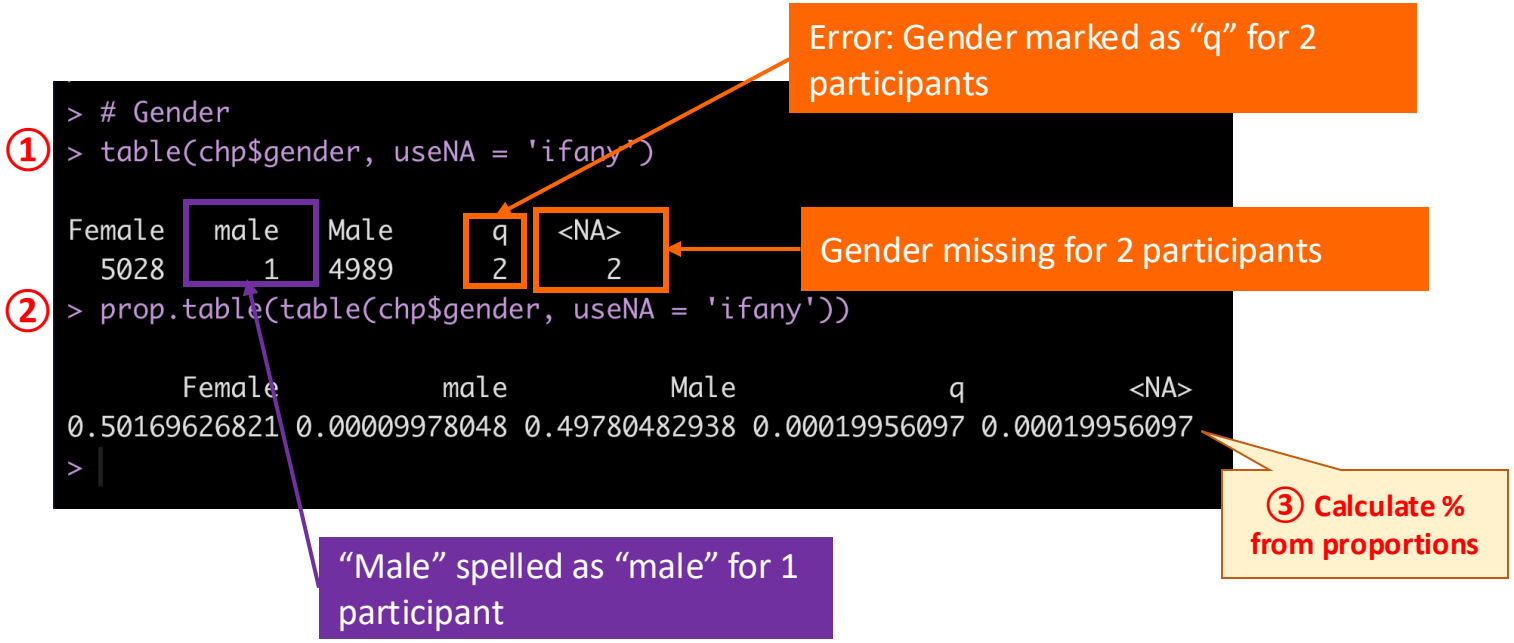
Summarising the data (categorical variables)

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.



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           zsmoker              <NA>
             1480              1120              6517               901                 2                 2
> 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
```

Error: smoke marked as "zsmoker" for 2 participants

Smoking status data missing for 2 participants

Ethnicity data missing for 3 participants

"Chinese" spelled as "chinese" for 1 participant

Data cleaning for continuous variable

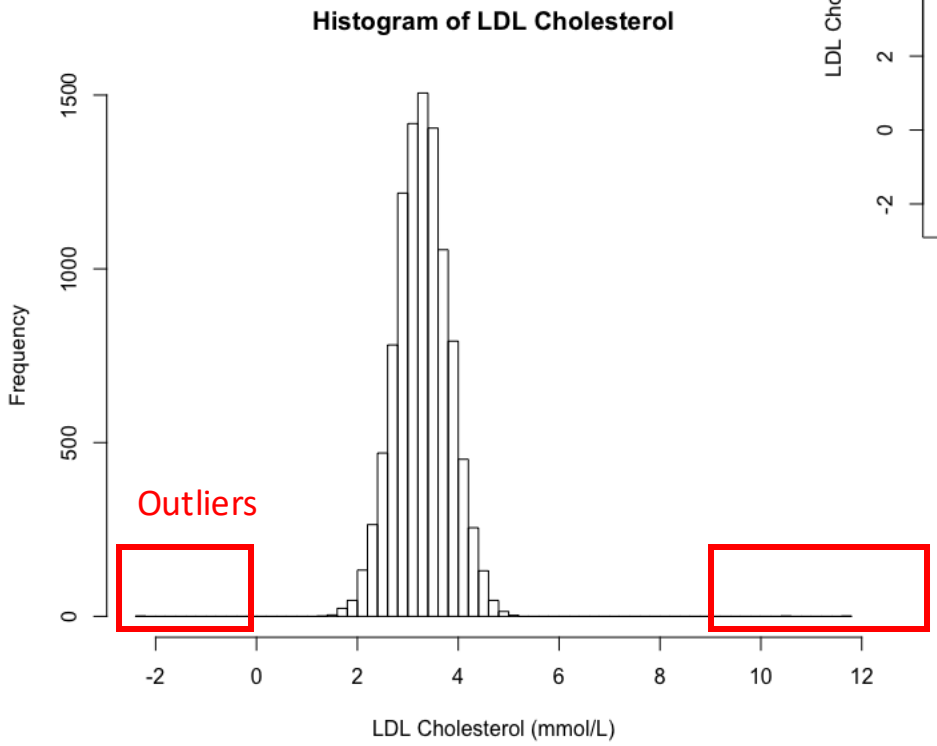
Data cleaning for continuous variable: LDL Cholesterol (recap)

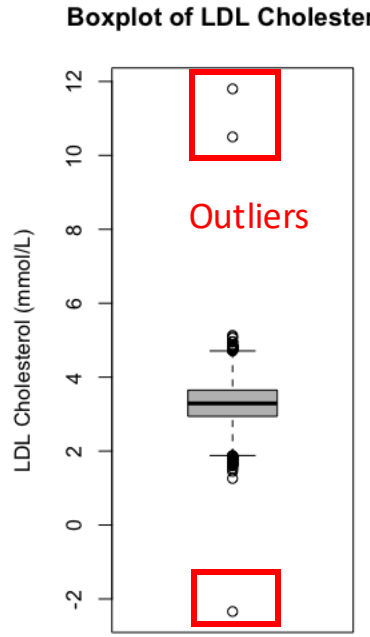
```
> summary(chp$ldl)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 -2.340   2.940   3.290   3.296   3.650  11.800       2
> quantile(chp$ldl,na.rm = TRUE)
   0%   25%   50%   75%  100%
-2.34  2.94  3.29  3.65 11.80
> # Checking the distribution of the variable (using histogram)
> hist(chp$ldl,breaks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

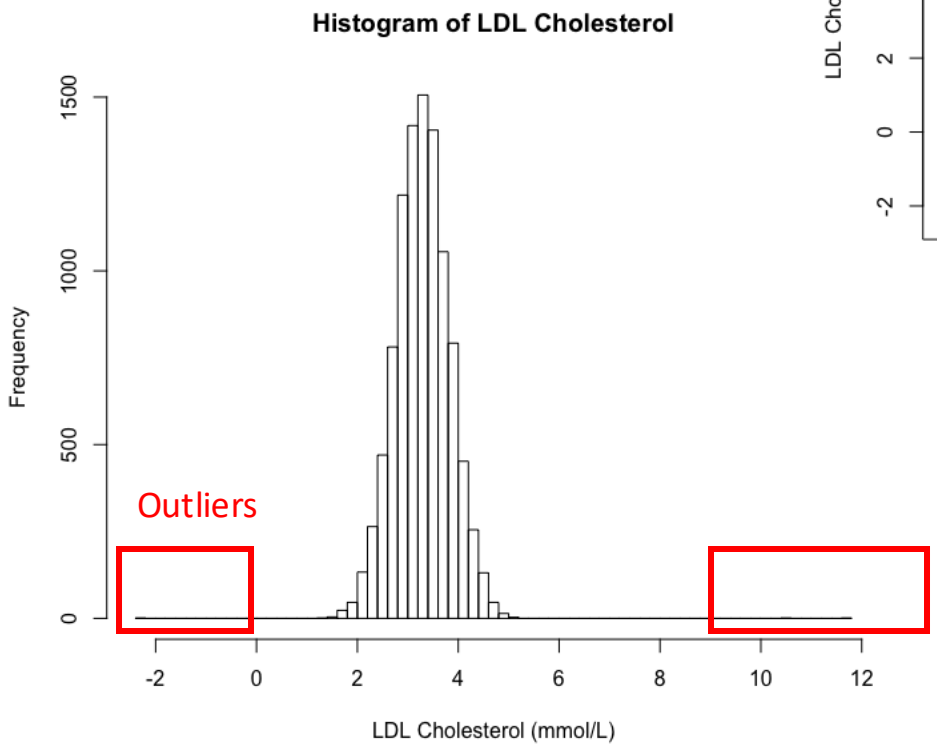
Erroneous data : negative values
LDL values missing for 2 participants

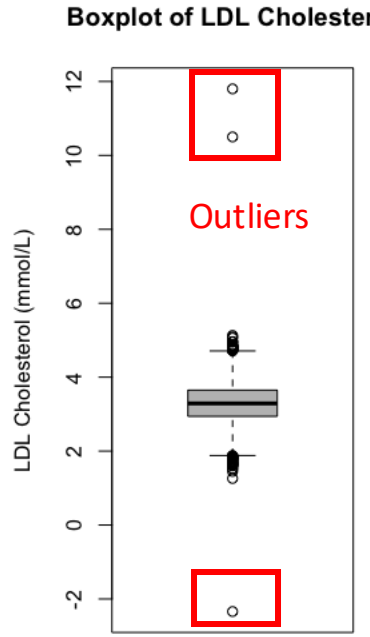
Some values appear to be too high

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

```
> # LDL
① > 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

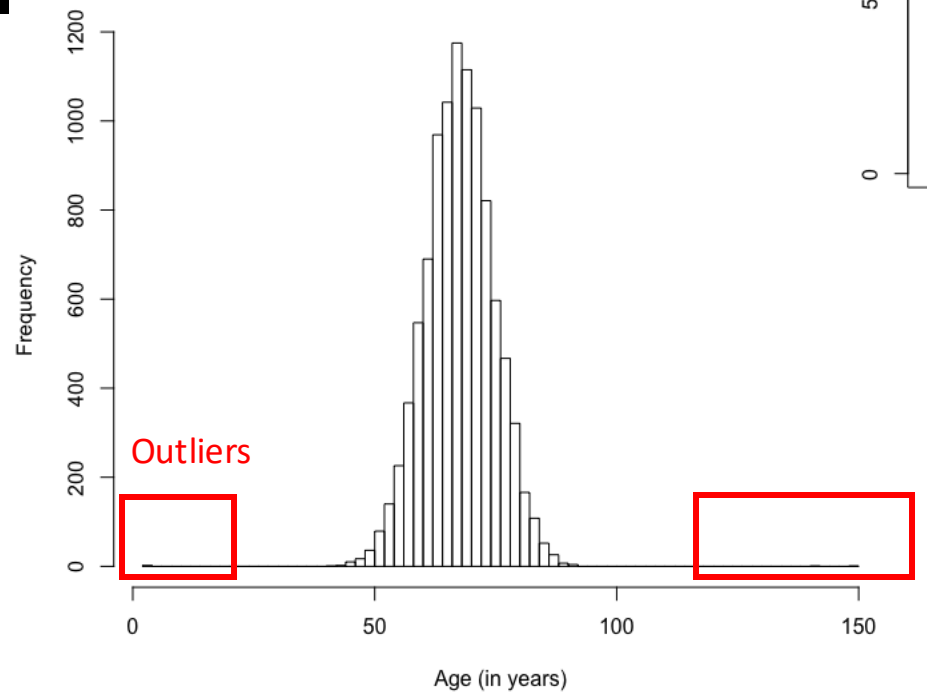
Age missing for some participants

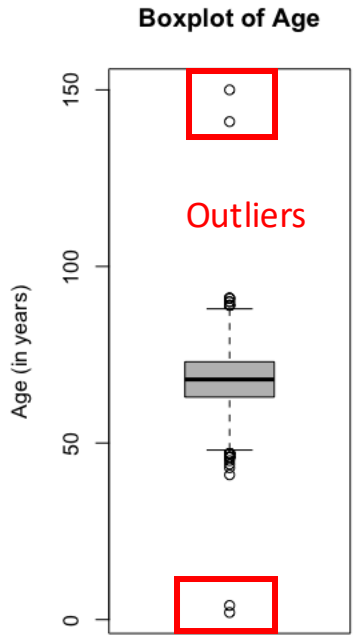
① Create new variable "age_drop" which has the same values as age (for now).
② For "age_drop" discard those outside the range of 40-100
③ Summarise the new variable

```
> # Age
> chp$age_drop = chp$age
> chp$age_drop[chp$age_drop<40 | chp$age_drop>100]
[1]   4 141 150   2  NA  NA  NA  NA
> chp$age_drop[chp$age_drop<40 | chp$age_drop>100] = NA
> 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
```

① ② ③

Verify the range





In-class Exercise #3 Solution

Some values appear to be too high

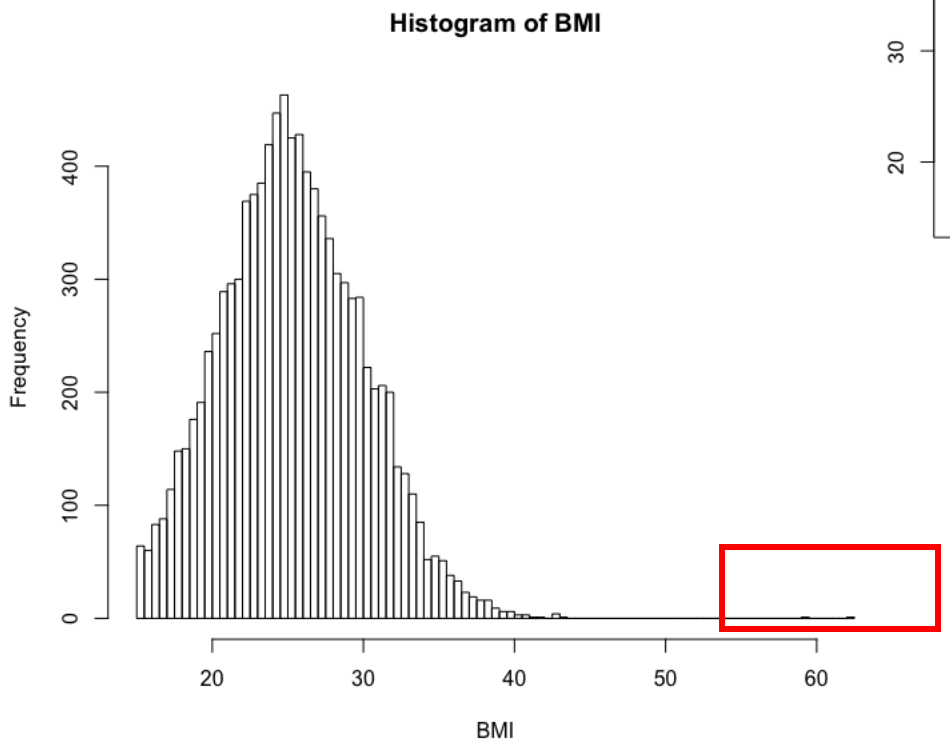
BMI missing for some participants

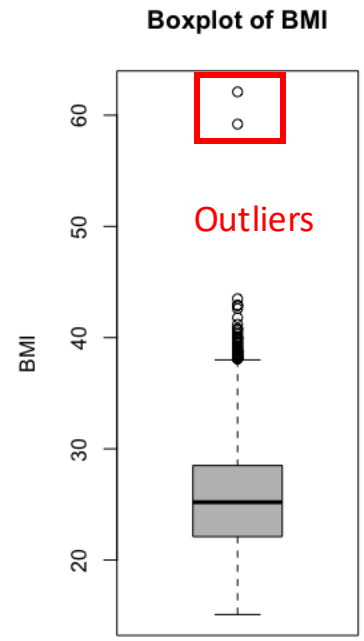
```
> # BMI
> summary(chp$bmi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  15.10   22.10   25.20   25.38   28.50   62.10       1
> quantile(chp$bmi,na.rm = TRUE)
  0%  25%  50%  75% 100%
15.1 22.1 25.2 28.5 62.1
> # Checking the distribution of the variable (using histogram)
> hist(chp$bmi,breaks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

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

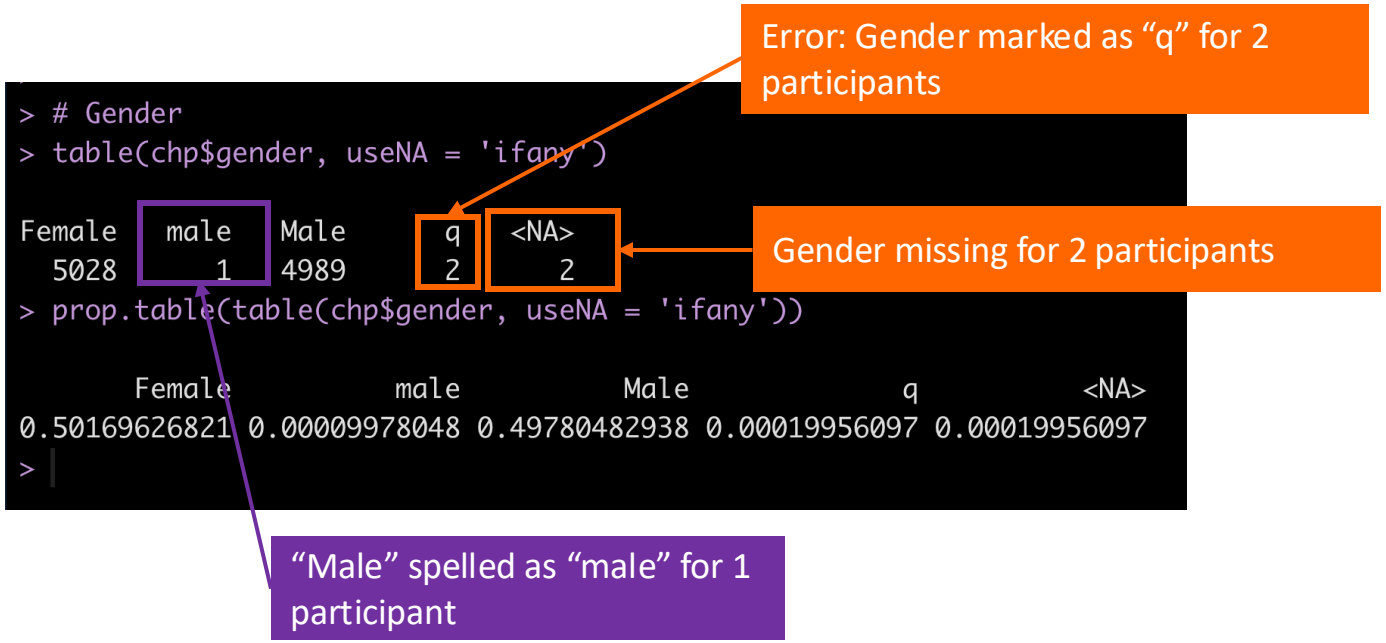
```
① > chp$bmi_drop = chp$bmi
  > chp$bmi_drop[chp$bmi_drop>50]
  [1]   NA 62.1 59.2
② > chp$bmi_drop[chp$bmi_drop>50] = NA
③ > summary(chp$bmi_drop)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  15.10   22.10   25.20   25.37   28.50   43.50       3
```

Verify the range





Data cleaning for categorical variable: gender (recap)



✓ Erroneous data; 2 observations have values “q”

✓ “Male” also spelled in lowercase “male

Data cleaning for categorical variable: gender

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

Female   male   Male      q   <NA>
  5028      1   4989      2      2
> prop.table(table(chp$gender, useNA = 'ifany'))

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

Error: Gender marked as "q" for 2 participants

Gender missing for 2 participants

"Male" spelled as "male" for 1 participant

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

③
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

Error: smoke marked as "zsmoker" for 2 participants

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

     Daily smoker          Ex-smoker      Never-Smoker Occasional smoker            zsmoker               <NA>
             1480               1120               6517                901                  2                  2
> 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               <NA>
             1480               1120               6517                901                  4
>
>
> # Ethnicity
> table(chp$ethnicity, useNA = 'ifany')

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

"Chinese" spelled as "chinese" for 1 participant

Verify the values ③

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

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    Male    <NA>
  5028    4990       4
> chp$female = NA
> table(chp$female, useNA = 'ifany')

 <NA>
10022
> chp$female[chp$gender %in% "Female"] = 1
> chp$female[chp$gender %in% "Male"] = 0
> table(chp$female, useNA = 'ifany')

   0    1 <NA>
4990 5028    4
```

Female coded as 1

Male coded as 0

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$smoke)
[1] "character"
> table(chp$smoke, useNA = 'ifany')

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

Converted from categorical to numeric

Smoking status has 4 unique values

Data transformation

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

```
> # Smoke
> class(chp$smoke)
[1] "character"
> table(chp$smoke, useNA = 'ifany')

     Daily smoker          Ex-smoker      Never-Smoker Occasional smoker              <NA>
             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 <NA>
⑥ 1480 1120 6517  901    4
```

Converted from categorical to numeric

Smoking status has 4 unique values

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
① > chp$smoker = NA
> table(chp$smoker, useNA = 'ifany')

 <NA>
10022
② > chp$smoker[chp$smoke %in% "Daily smoker"] = 1
③ > chp$smoker[chp$smoke %in% "Occasional smoker"] = 1
④ > chp$smoker[chp$smoke %in% "Ex-smoker"] = 0
⑤ > chp$smoker[chp$smoke %in% "Never-Smoker"] = 0
> 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
```

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

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

Data transformation : Categorising a continuous variable

Categorise variable bmi into binary variable [Overweight (Yes: ≥25 kg/m^2 ; No: ≤24.9 kg/m^2)]

```
> # Re-categorise variable bmi into binary variable [overweight (1: ≥25 kg/m2 ; 0: ≤24.9 kg/m2)]
> 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 kg/m^2
③ Code “overweight” as 1 if BMI ≥ 25 kg/m^2
④ Verify the range

Data transformation : Categorising a continuous variable

Categorise variable bmi into binary variable [Overweight (Yes: ≥25 kg/m² ; No: ≤24.9 kg/m²)]

```
> # Re-categorise variable bmi into binary variable [overweight (1: ≥25 kg/m2 ; 0: ≤24.9 kg/m2)]
① > 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
```

Check the range for each group

① Create new variable “overweight” using the $ operator
② Code “overweight” as 0 if BMI < 25 kg/m²
③ Code “overweight” as 1 if BMI ≥ 25 kg/m²
④ Verify the range

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

```
> # Identify the rows with missing and/or erroneous data using the "which" function
① > 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_dro
p))
> 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
② > 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