Show me the data!

Week02: R, Vector, and Object

Big Data & Social Analysis R

Instructors： Chung-pei Pien

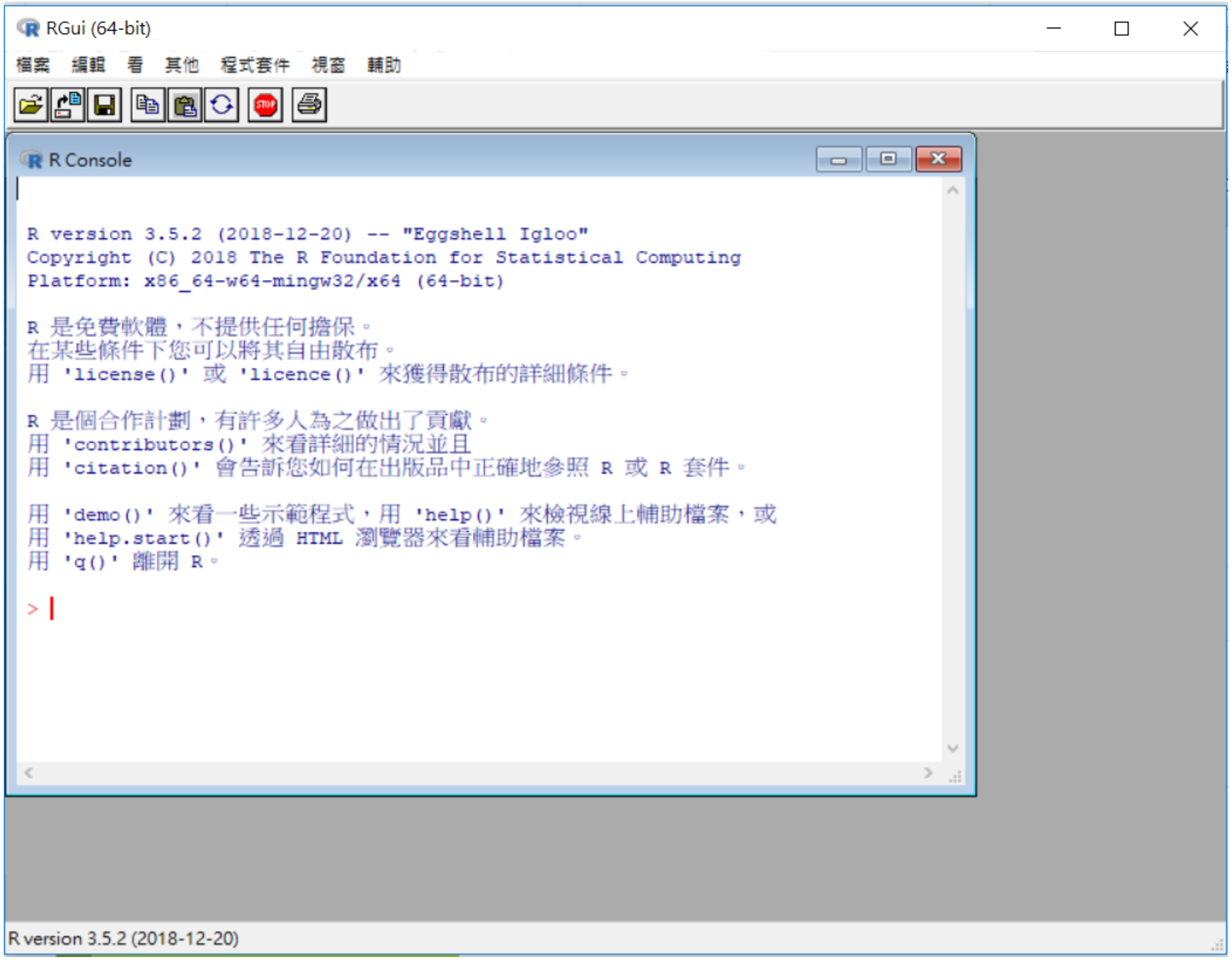
ZU1942001/266868001/Z23937001/ZM1941001

International College of INNOVATION
National Chengchi University
國立政治大學創新國際學院

CONTENTS

R and R-Studio

01 R and R-Studio


RGui (64-bit)
檔案 編輯 看 其他 程式套件 視窗 輔助
R Console
R version 3.5.2 (2018-12-20)

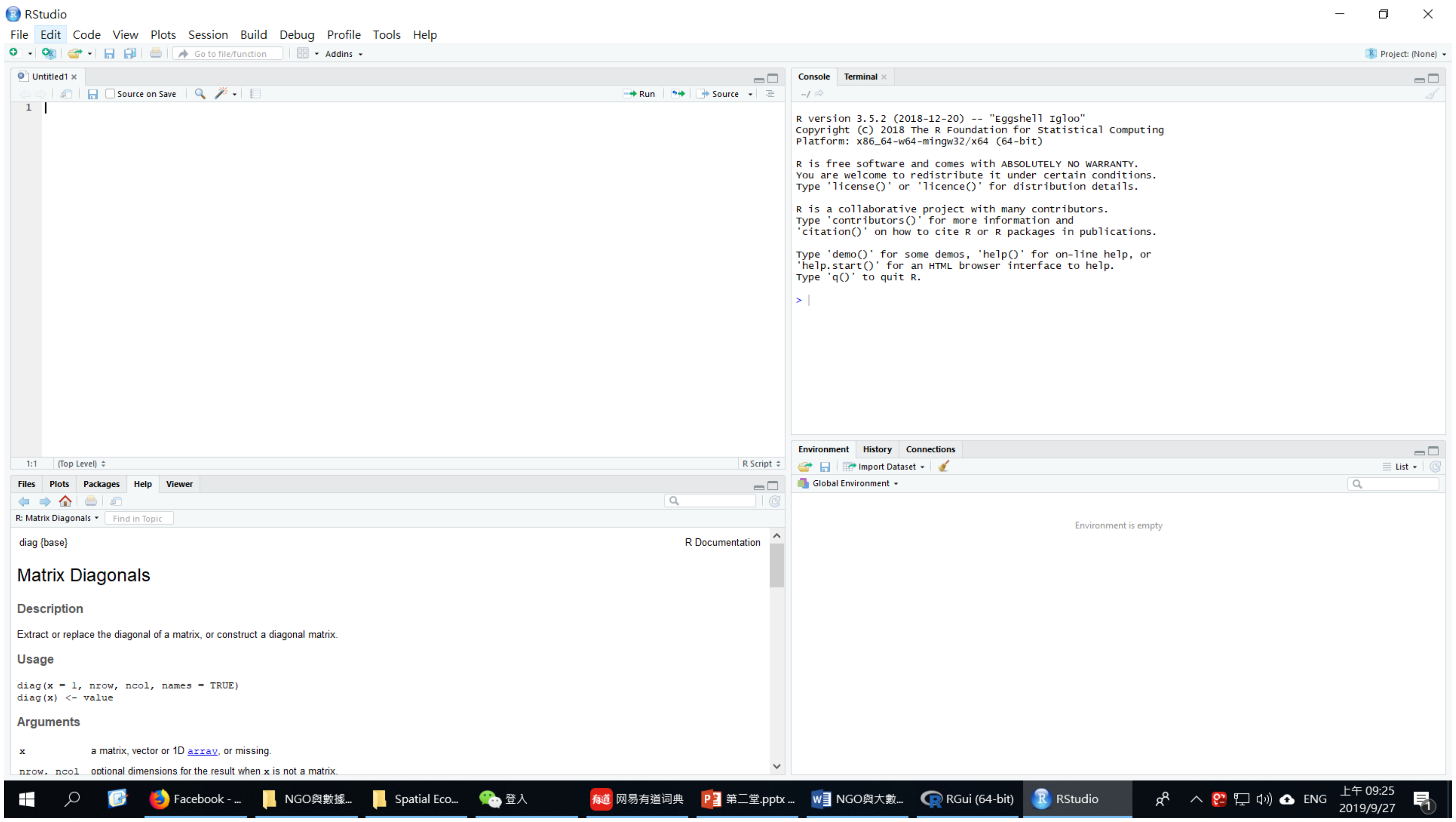
```
R version 3.5.2 (2018-12-20) -- "Eggshell Igloo"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R 是免費軟體，不提供任何擔保。
在某些條件下您可以將其自由散布。
用 'license()' 或 'licence()' 來獲得散布的詳細條件。

R 是個合作計劃，有許多人為之做出了貢獻。
用 'contributors()' 來看詳細的情況並且
用 'citation()' 會告訴您如何在出版品中正確地參照 R 或 R 套件。

用 'demo()' 來看一些示範程式，用 'help()' 來檢視線上輔助檔案，或
用 'help.start()' 透過 HTML 瀏覽器來看輔助檔案。
用 'q()' 離開 R。

>
```

01

R and R-Studio

R： Base computer

R-Studio： add RAMs, two monitors



```
R version 3.5.2 (2018-12-20) -- "Eggshell Igloo"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (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.

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.

>
```

diag {base} R Documentation

Matrix Diagonals

Description

Extract or replace the diagonal of a matrix, or construct a diagonal matrix.

Usage

```
diag(x = 1, nrow, ncol, names = TRUE)
diag(x) <- value
```

Arguments

`x`	a matrix, vector or 1D array, or missing.
`nrow, ncol`	optional dimensions for the result when x is not a matrix.

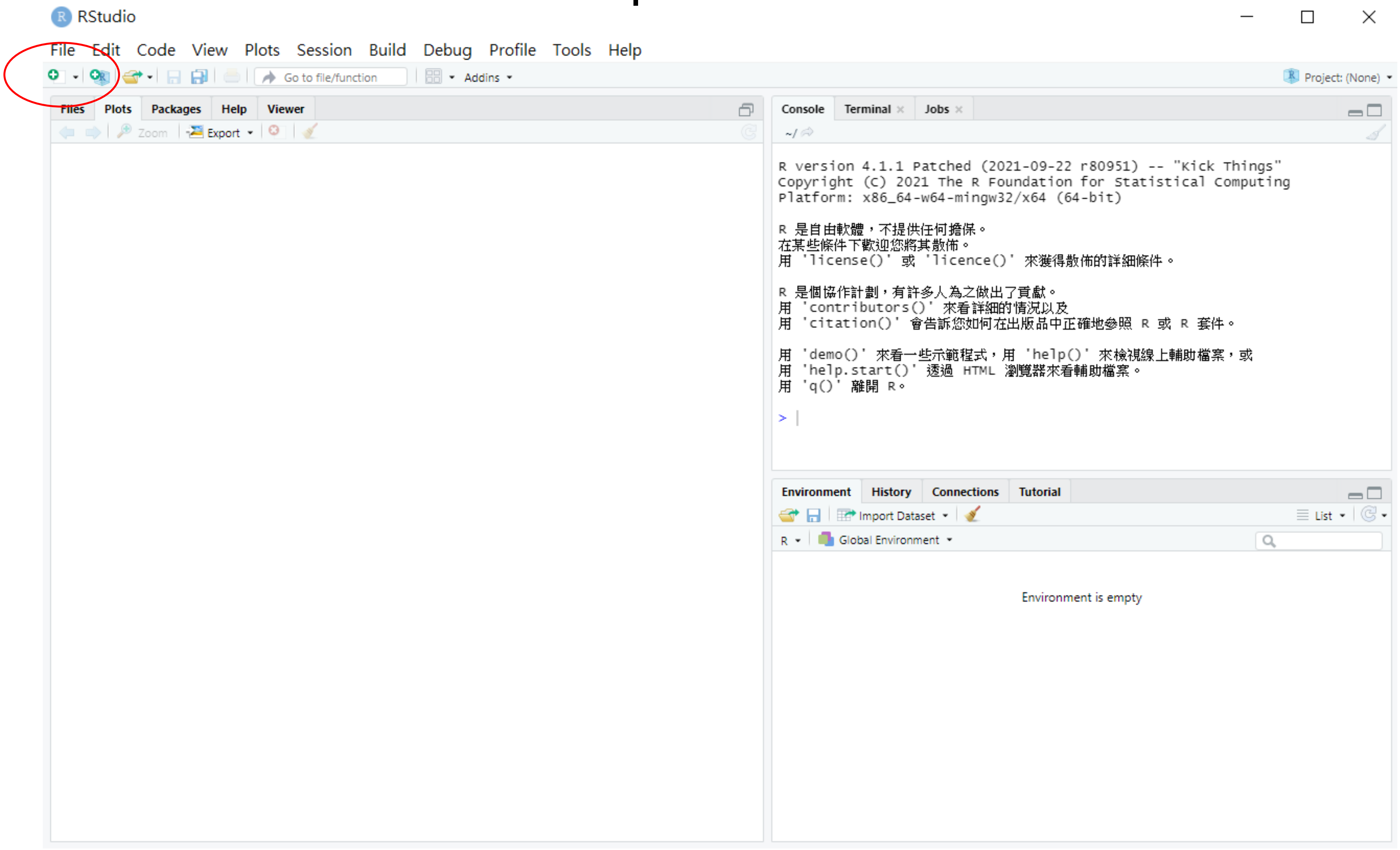
01 R and R-Studio

- Console：Input commands and show results
- Environment：the list of objects
- Source：R scripts and content of objects
- Plots：the plots
- Packages：the list of packages
- Help

01

R and R-Studio

Create a new R script file

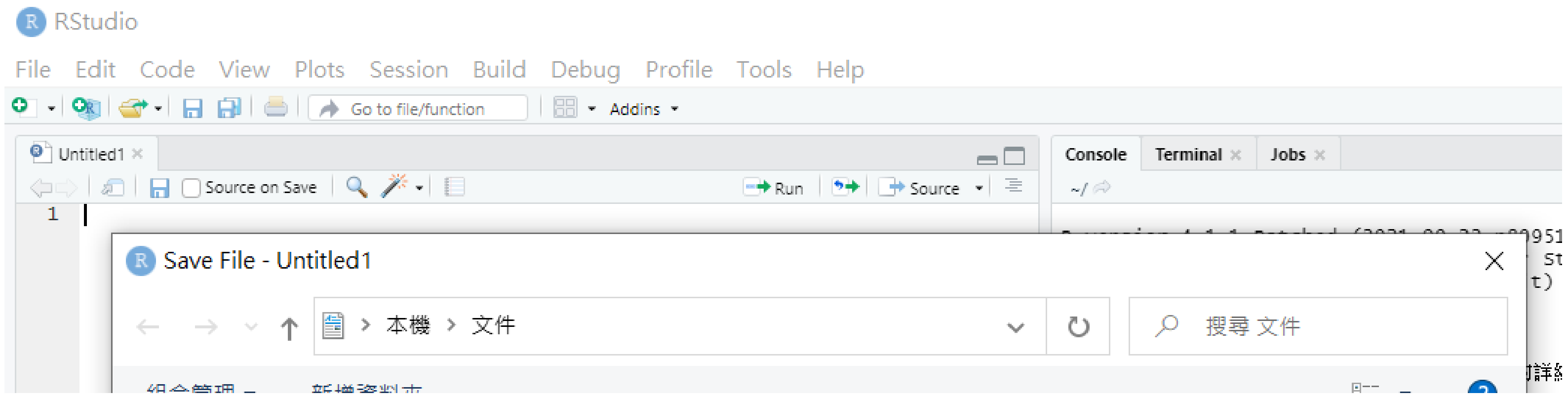


```
R version 4.1.1 Patched (2021-09-22 r80951) -- "Kick Things"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R 是自由軟體，不提供任何擔保。
在某些條件下歡迎您將其散佈。
用 'license()' 或 'licence()' 來獲得散佈的詳細條件。

R 是個協作計劃，有許多人為之做出了貢獻。
用 'contributors()' 來看詳細的情況以及
用 'citation()' 會告訴您如何在出版品中正確地參照 R 或 R 套件。

用 'demo()' 來看一些示範程式，用 'help()' 來檢視線上輔助檔案，或
用 'help.start()' 透過 HTML 瀏覽器來看輔助檔案。
用 'q()' 離開 R。

> |
```

01

R and R-Studio

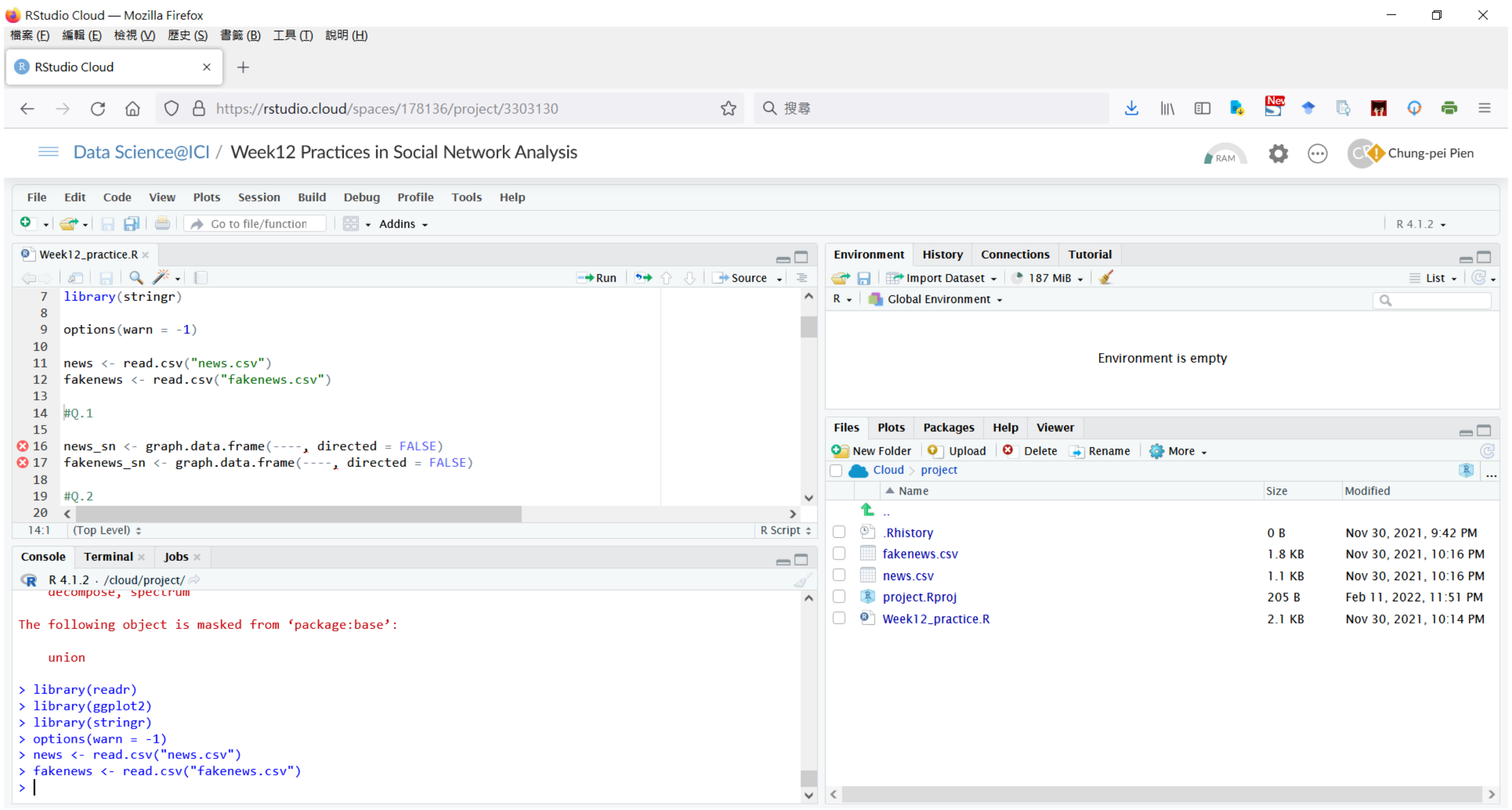
1. Create a new project folder
2. Save the R script file into this folder



01

R and R-Studio

R-Studio Cloud: A cloud-based solution that allows anyone to do, share, teach and learn data science online


RStudio Cloud — Mozilla Firefox
檔案 (F) 編輯 (E) 檢視 (V) 歷史 (S) 書籤 (B) 工具 (T) 說明 (H)
RStudio Cloud
https://rstudio.cloud/spaces/178136/project/3303130
搜尋
Data Science@ICI / Week12 Practices in Social Network Analysis
RAM
Chung-pei Pien
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.1.2
Week12_practice.R
Run
Source
14:1
(Top Level)
R Script
Console Terminal Jobs
R 4.1.2 · /cloud/project/
Environment History Connections Tutorial
Import Dataset
187 MiB
List
R
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder Upload Delete Rename More
Cloud › project

```r
library(stringr)

options(warn = -1)

news <- read.csv("news.csv")
fakenews <- read.csv("fakenews.csv")

#Q.1

news_sn <- graph.data.frame(----, directed = FALSE)
fakenews_sn <- graph.data.frame(----, directed = FALSE)

#Q.2
```

```
    decompose, spectrum

The following object is masked from ‘package:base’:

    union

> library(readr)
> library(ggplot2)
> library(stringr)
> options(warn = -1)
> news <- read.csv("news.csv")
> fakenews <- read.csv("fakenews.csv")
> |
```

Name	Size	Modified
..		
.Rhistory	0 B	Nov 30, 2021, 9:42 PM
fakenews.csv	1.8 KB	Nov 30, 2021, 10:16 PM
news.csv	1.1 KB	Nov 30, 2021, 10:16 PM
project.Rproj	205 B	Feb 11, 2022, 11:51 PM
Week12_practice.R	2.1 KB	Nov 30, 2021, 10:14 PM

Basic Skills of R

The basic concept of R:

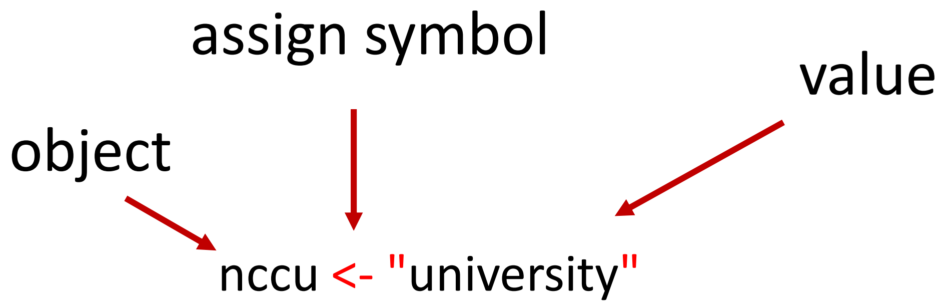
Assign values to objects

The basic concept of coding in R is to assign a set of values to objects. Then you can transform, analyze, calculate, or represent the objects

Basic Skills of R

Object and Assign

Assign values to an object

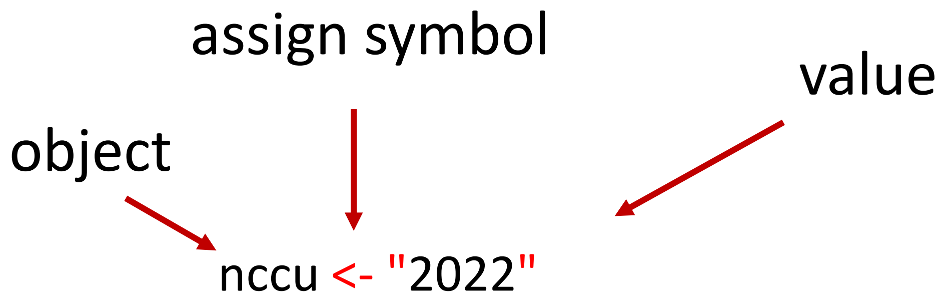


```
nccu <- "university"
```

Basic Skills of R

Object and Assign

Assign values to an object



```
nccu <- "2022"
```

Basic Skills of R

Introduction of Object

Objects is the key element in R：

1. An object can involve one or million data.

2. There are different kinds of objects. They offer different ways to store data.

02

Basic Skills of R

Introduction of Object

The kinds of objects in R

1. Vector
2. List
3. Matrix
4. Table (dataframe)
5. Others:

Basic Skills of R

Introduction of Object

The kinds of objects in R

1. Vector
2. List
3. Matrix
4. Table (dataframe)
5. Others:

02

Basic Skills of R

Vector

Vector: a set of numbers (numeric) or characters is created by c() function.

```
nccu <- c("taipei", "ici", "zoo", "beautiful")

z <- c(1, 5, 10, 15, 27)
```

Basic Skills of R

Vector

If a vector includes numerics and or characters?

```
nccu <- c("taipei", "ici", "zoo", "beautiful", 1, 6, 12)
```

R recognizes it as Character

02

Basic Skills of R

Vector

To create/identify an object as numerics or characters is extremely important in R!!!

```
lawmaker_id <- c("0123", "1123", "0894”, "1305")
```

```
lawmaker_id2 <- c(0123, 1123, 0894, 1305)
```

Basic Skills of R

Vector

What is coding?

Coding is to pick up specific elements in your objects and calculate/change their value.

```
nccu <- c("taipei", "ici", "zoo", "beautiful")

z <- c(1, 5, 10, 15, 27)
```

02

Basic Skills of R

Vector

What is coding?

Coding is to pick up specific elements in your objects and calculate/change their value.

```
nccu <- c("taipei", "ici", "zoo", "beautiful")

nccu[1]
nccu[3]
nccu[5]
```

02

Basic Skills of R

Vector

What is coding?

Coding is to pick up specific elements in your objects and calculate/change their value.

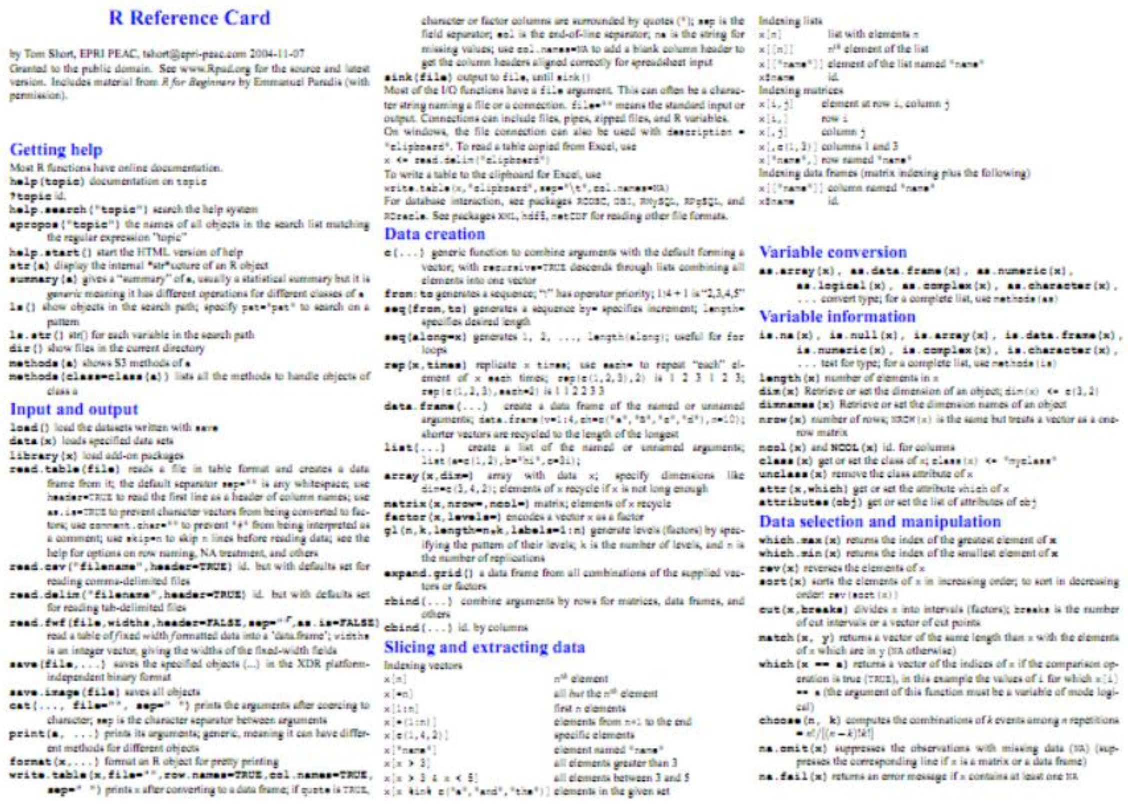
```
nccu <- c("taipei", "ici", "zoo", "beautiful")

nccu[1]
nccu[3]
nccu[5]

nccu[1] <- c("home")

nccu

nccu[5] <- c("international")
```

Basic Skills of R

02

Function

A Function in R is to do a specific task, such as creating objects, picking up elememnts, making plots, creating models, and so on.



R Reference Card

by Tom Short, EPRI PEAC, tshort@epri-peac.com 2004-11-07
Granted to the public domain. See www.Rpad.org for the source and latest version. Includes material from *R for Beginners* by Emmanuel Paradis (with permission).

Getting help

Most R functions have online documentation.

help(topic) documentation on topic
?topic id.
help.search("topic") search the help system
apropos("topic") the names of all objects in the search list matching the regular expression "topic"
help.start() start the HTML version of help
str(a) display the internal *str*ucture of an R object
summary(a) gives a "summary" of a, usually a statistical summary but it is *generic* meaning it has different operations for different classes of a
ls() show objects in the search path; specify pat="pat" to search on a pattern
ls.str() str() for each variable in the search path
dir() show files in the current directory
methods(a) shows S3 methods of a
methods(class=class(a)) lists all the methods to handle objects of class a

Input and output

load() load the datasets written with save
data(x) loads specified data sets
library(x) load add-on packages
read.table(file) reads a file in table format and creates a data frame from it; the default separator sep="" is any whitespace; use header=TRUE to read the first line as a header of column names; use as.is=TRUE to prevent character vectors from being converted to factors; use comment.char="" to prevent "#" from being interpreted as a comment; use skip=n to skip n lines before reading data; see the help for options on row naming, NA treatment, and others
read.csv("filename",header=TRUE) id. but with defaults set for reading comma-delimited files
read.delim("filename",header=TRUE) id. but with defaults set for reading tab-delimited files
read.fwf(file,widths,header=FALSE,sep="",as.is=FALSE) read a table of *f*ixed *w*idth *f*ormatted data into a 'data.frame'; widths is an integer vector, giving the widths of the fixed-width fields
save(file,...) saves the specified objects (...) in the XDR platform-independent binary format
save.image(file) saves all objects
cat(..., file="", sep=" ") prints the arguments after coercing to character; sep is the character separator between arguments
print(a, ...) prints its arguments; generic, meaning it can have different methods for different objects
format(x,...) format an R object for pretty printing
write.table(x,file="",row.names=TRUE,col.names=TRUE, sep=" ") prints x after converting to a data frame; if quote is TRUE, character or factor columns are surrounded by quotes ("); sep is the field separator; eol is the end-of-line separator; na is the string for missing values; use col.names=NA to add a blank column header to get the column headers aligned correctly for spreadsheet input
sink(file) output to file, until sink()

Most of the I/O functions have a file argument. This can often be a character string naming a file or a connection. file="" means the standard input or output. Connections can include files, pipes, zipped files, and R variables.
On windows, the file connection can also be used with description = "clipboard". To read a table copied from Excel, use
x <- read.delim("clipboard")
To write a table to the clipboard for Excel, use
write.table(x,"clipboard",sep="\t",col.names=NA)
For database interaction, see packages RODBC, DBI, RMySQL, RPgSQL, and ROracle. See packages XML, hdf5, netCDF for reading other file formats.

Data creation

c(...) generic function to combine arguments with the default forming a vector; with recursive=TRUE descends through lists combining all elements into one vector
from:to generates a sequence; ":" has operator priority; 1:4 + 1 is "2,3,4,5"
seq(from,to) generates a sequence by= specifies increment; length= specifies desired length
seq(along=x) generates 1, 2, ..., length(along); useful for for loops
rep(x,times) replicate x times; use each= to repeat "each" element of x each times; rep(c(1,2,3),2) is 1 2 3 1 2 3; rep(c(1,2,3),each=2) is 1 1 2 2 3 3
data.frame(...) create a data frame of the named or unnamed arguments; data.frame(v=1:4,ch=c("a","B","c","d"),n=10); shorter vectors are recycled to the length of the longest
list(...) create a list of the named or unnamed arguments; list(a=c(1,2),b="hi",c=3i);
array(x,dim=) array with data x; specify dimensions like dim=c(3,4,2); elements of x recycle if x is not long enough
matrix(x,nrow=,ncol=) matrix; elements of x recycle
factor(x,levels=) encodes a vector x as a factor
gl(n,k,length=n*k,labels=1:n) generate levels (factors) by specifying the pattern of their levels; k is the number of levels, and n is the number of replications
expand.grid() a data frame from all combinations of the supplied vectors or factors
rbind(...) combine arguments by rows for matrices, data frames, and others
cbind(...) id. by columns

Slicing and extracting data

Indexing vectors

x[n]	n^{th} element
x[-n]	all *but* the n^{th} element
x[1:n]	first *n* elements
x[-(1:n)]	elements from n+1 to the end
x[c(1,4,2)]	specific elements
x["name"]	element named "name"
x[x > 3]	all elements greater than 3
x[x > 3 & x < 5]	all elements between 3 and 5
x[x %in% c("a","and","the")]	elements in the given set

Indexing lists

x[n]	list with elements n
x[[n]]	n^{th} element of the list
x[["name"]]	element of the list named "name"
x$name	id.

Indexing matrices

x[i,j]	element at row i, column j
x[i,]	row i
x[,j]	column j
x[,c(1,3)]	columns 1 and 3
x["name",]	row named "name"

Indexing data frames (matrix indexing plus the following)

x[["name"]]	column named "name"
x$name	id.

Variable conversion

as.array(x), as.data.frame(x), as.numeric(x), as.logical(x), as.complex(x), as.character(x), ... convert type; for a complete list, use methods(as)

Variable information

is.na(x), is.null(x), is.array(x), is.data.frame(x), is.numeric(x), is.complex(x), is.character(x), ... test for type; for a complete list, use methods(is)
length(x) number of elements in x
dim(x) Retrieve or set the dimension of an object; dim(x) <- c(3,2)
dimnames(x) Retrieve or set the dimension names of an object
nrow(x) number of rows; NROW(x) is the same but treats a vector as a one-row matrix
ncol(x) and **NCOL(x)** id. for columns
class(x) get or set the class of x; class(x) <- "myclass"
unclass(x) remove the class attribute of x
attr(x,which) get or set the attribute which of x
attributes(obj) get or set the list of attributes of obj

Data selection and manipulation

which.max(x) returns the index of the greatest element of x
which.min(x) returns the index of the smallest element of x
rev(x) reverses the elements of x
sort(x) sorts the elements of x in increasing order; to sort in decreasing order: rev(sort(x))
cut(x,breaks) divides x into intervals (factors); breaks is the number of cut intervals or a vector of cut points
match(x, y) returns a vector of the same length than x with the elements of x which are in y (NA otherwise)
which(x == a) returns a vector of the indices of x if the comparison operation is true (TRUE), in this example the values of i for which x[i] == a (the argument of this function must be a variable of mode logical)
choose(n, k) computes the combinations of *k* events among *n* repetitions $= n!/[(n-k)!k!]$
na.omit(x) suppresses the observations with missing data (NA) (suppresses the corresponding line if x is a matrix or a data frame)
na.fail(x) returns an error message if x contains at least one NA

02

Basic Skills of R

Vector Example 1

There is a vector student to show a class’s students’ gender information.

```
student <- c("m", "f", "f", "m", "m", "m", "f", "f", "f", "f")
```

1. How many students are in this class?

2. How many male students are in this class?

3. How many female students are in this class?

02

Basic Skills of R

Vector Example 1

There is a vector student to show a class’s students’ gender information.

```
student <- c("m", "f", "f", "m", "m", "m", "f", "f", "f", "f")
```

1. How many students are in this class?

```
length(student)
```

2. How many male students are in this class?

```
student[student == "m"]
length(student[student == "m"])
```

02

Basic Skills of R

Vector Example 1

There is a vector student to show a class's students' gender information.

```
student <- c("m", "f", "f", "m", "m", "m", "f", "f", "f", "f")
```

3. How many female students are in this class?

Answer this question in Moodle (Practice 1)

02

Basic Skills of R

Vector Example 2

There is a vector age to show a company’s workers’ age.

```
age <- c(45, 60, 22, 61, 34, 59, 64, 54, 29, 31)
```

1. How many workers are in this company?
2. How many workers’ age are larger than 60?
3. How many workers’ age are smaller than 30?

02

Basic Skills of R

Vector Example 2

There is a vector age to show a company’s workers’ age.

```
age <- c(45, 60, 22, 61, 34, 59, 64, 54, 29, 31)
```

1. How many workers are in this company?

```
length(age)
```

2. How many workers’ age are larger than 60?

```
age[age >= 60]
length(age[age >= 60])
```

02

Basic Skills of R

Vector Example 2

There is a vector age to show a company’s workers’ age.

```
age <- c(45, 60, 22, 61, 34, 59, 64, 54, 29, 31)
```

3. How many workers’ age are smaller than 30?

Answer this question in Moodle (Practice 2)

02

Basic Skills of R

Vector Example 1

There is a vector student to show a class’s students’ gender information.

```
student <- c("m", "f", "f", "m", "m", "m", "f", "f", "f", "f")
```

4. The ratio of male students in this class

```
    length(student[student == "m"]) / length(student)
```

02

Basic Skills of R

Vector Example 3

There is a vector id to show a NCCU class’s undergraduate students’ university id.

```
id <- c("110111222", "109222111", "109222333", "108333444",
"110555666")
```

1. How many freshman students are in this class?

02

Basic Skills of R

Vector Example 3

There is a vector id to show a NCCU class’s undergraduate students’ university id.

```
id <- c("110111222", "109222111", "109222333", "108333444",
"110555666")
```

1. How many freshman students are in this class?

```
nchar(id)
year <- substr(id, 1, 3)
length(year[year == "110"])
```

02

Basic Skills of R

Vector Example 4

There is a vector uid to show a NCCU class's students' university id.

```
uid <- c("110101222", "109252111", "109202333", "108351444",
"110151666")
```

1. How many undergraduate students are in this class?
2. How many undergraduate and freshman students are in this class?

Answer these questions in Moodle (Practice 3)

Basic Skills of R

Vector Example 5

There is a vector birth to show a NCCU class's students' birth year.

```
birth <- c("2002", "1999", "2001", "1998", "2000")
```

1. What is the average of the students' age

Basic Skills of R

Vector Example 5

There is a vector birth to show a NCCU class's students' birth year.

```
birth <- c("2002", "1999", "2001", "1998", "2000")
```

1. What is the average of the students' age

```
birth <- as.numeric(birth)
age_b <- 2022 - birth
mean(age_b)

median(age_b)
max(age_b)
```

Basic Skills of R

Vector Example 6

There is a vector birthdate to show a baseball team players’ birth date.

birthdate <- c("1983-11-01", "1995-01-19", "2001-06-23", "1987-12-09", "1999-10-21", "1999-03-31")

1. What is the average of the players’ age

Answer this question in Moodle (Practice 4)

115th US Congress Data

03

115th US Congress Data

Introduction

The term of office of 115th US Congress was from January 3, 2017, to January 3, 2019.

It was reelected in November 2018, two years after Trump's win in 2016.

The best form to use this data is a table. We will do it in the next week. Today, we still use vectors to handle this data.

115th US Congress Data

Project

Gender, year, party are three vectors to represent the 115th US House lawmakers' characteristics.

Please answer the following question in Moodle (Practice 5):

1. The female ratio of the lawmakers
2. The mean of lawmakers' age
3. Which party dominate the House?

Before you conduct this project, please think about why lawmakers' gender ratio, mean of age, and party domination are important?

Assignment

04

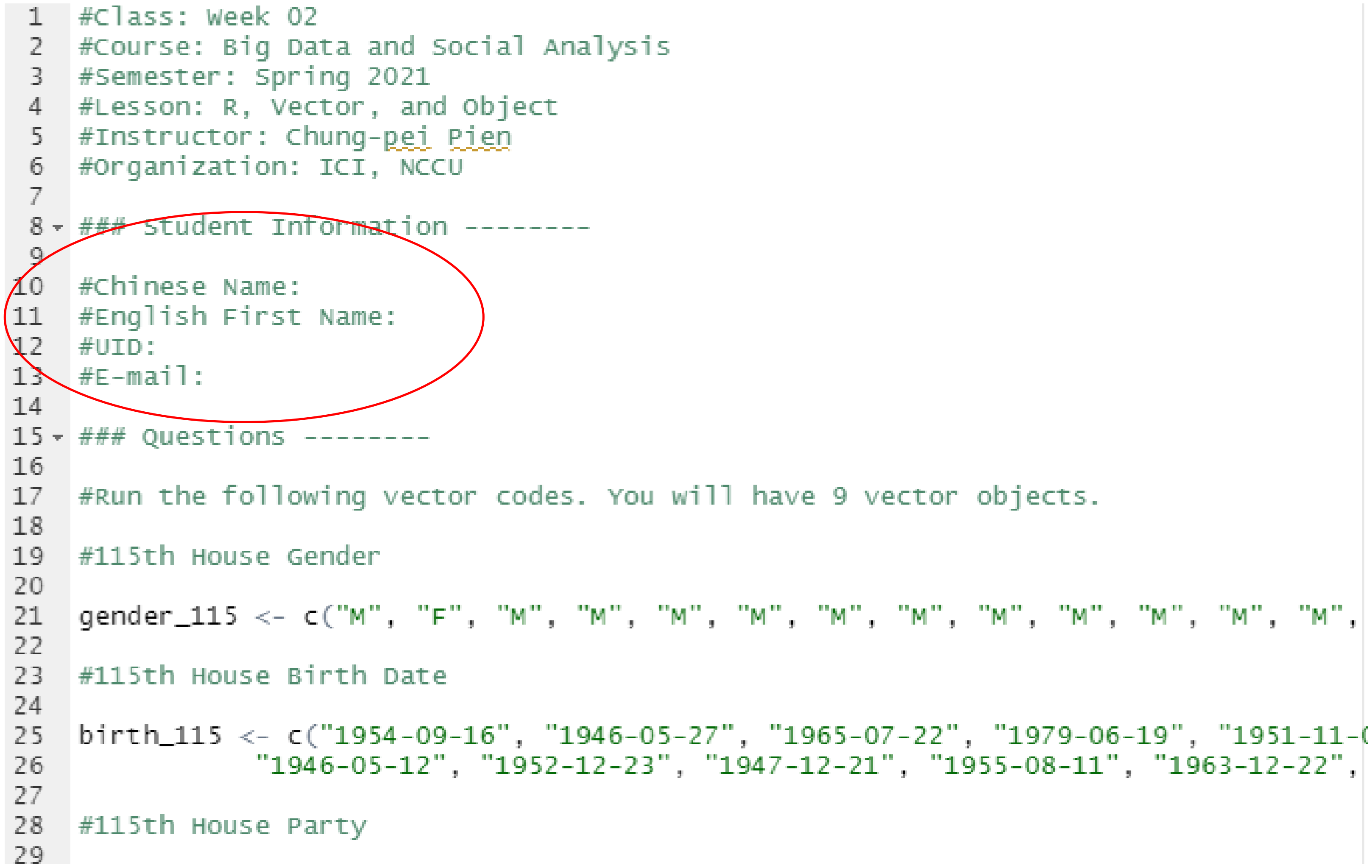
Assignment

115th -117th US Congress Data

This week's homework will answer the questions about the changes in gender inequality, age, party's control from 115th to 117th US congress.

04 Assignment

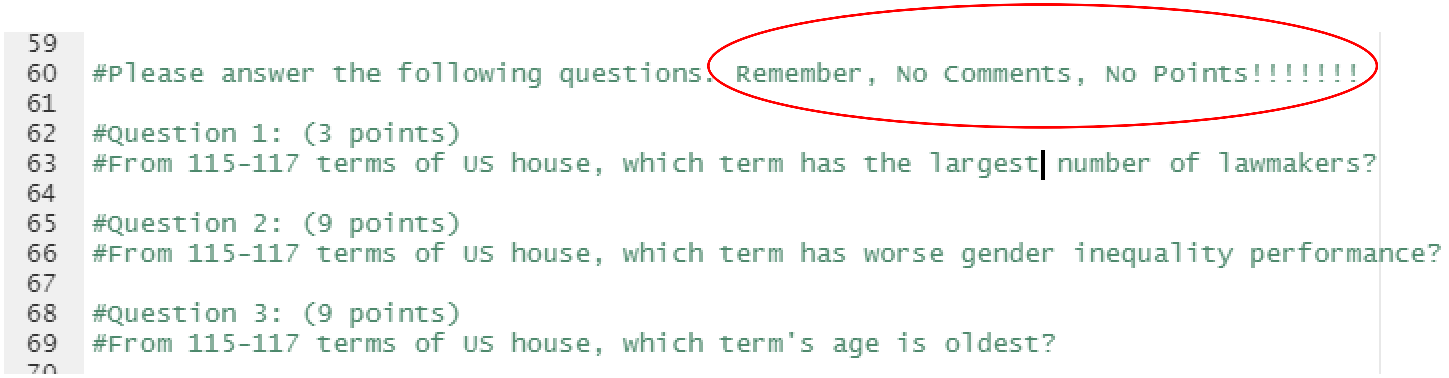
115th -117th US Congress Data



```
#Class: Week 02
#Course: Big Data and Social Analysis
#Semester: Spring 2021
#Lesson: R, Vector, and Object
#Instructor: Chung-pei Pien
#Organization: ICI, NCCU

### Student Information --------

#Chinese Name:
#English First Name:
#UID:
#E-mail:

### Questions --------

#Run the following vector codes. You will have 9 vector objects.

#115th House Gender

gender_115 <- c("M", "F", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M",

#115th House Birth Date

birth_115 <- c("1954-09-16", "1946-05-27", "1965-07-22", "1979-06-19", "1951-11-
           "1946-05-12", "1952-12-23", "1947-12-21", "1955-08-11", "1963-12-22",

#115th House Party

```

04 Assignment

115th -117th US Congress Data

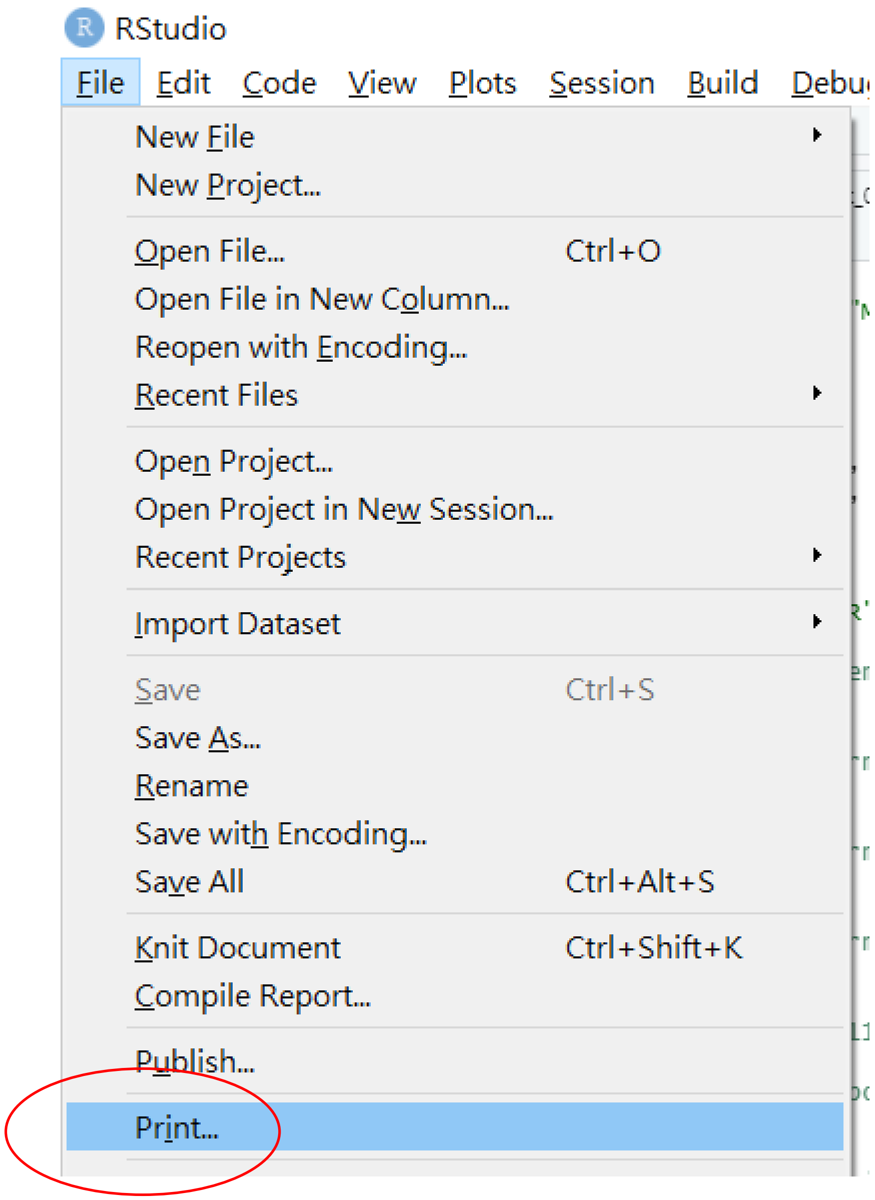


```

#Please answer the following questions. Remember, No Comments, No Points!!!!!!!

#Question 1: (3 points)
#From 115-117 terms of US house, which term has the largest number of lawmakers?

#Question 2: (9 points)
#From 115-117 terms of US house, which term has worse gender inequality performance?

#Question 3: (9 points)
#From 115-117 terms of US house, which term's age is oldest?
```

04

Assignment

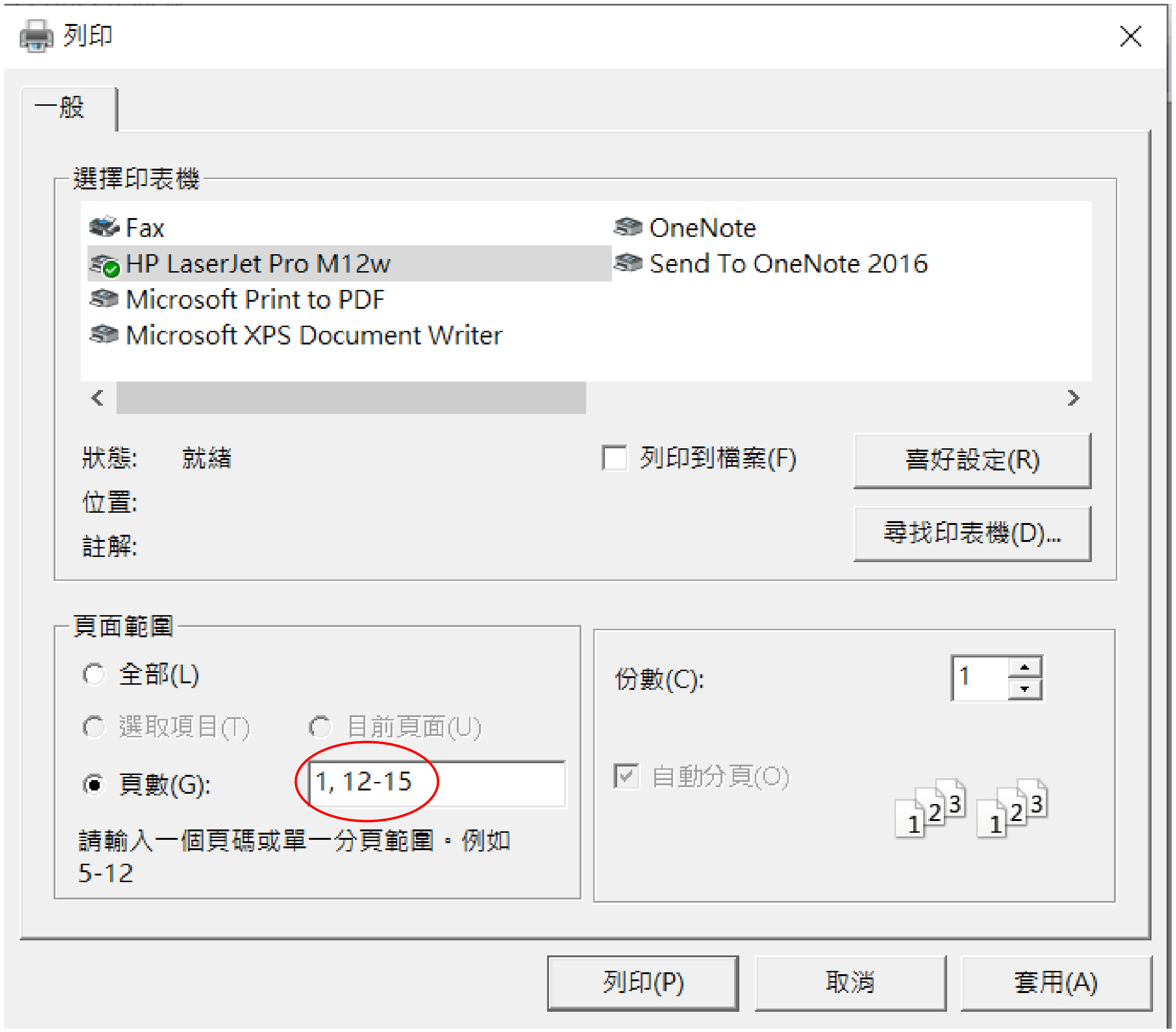
115th -117th US Congress Data



04

Assignment

115th -117th US Congress Data


列印

一般

選擇印表機

Fax
HP LaserJet Pro M12w
Microsoft Print to PDF
Microsoft XPS Document Writer
OneNote
Send To OneNote 2016

狀態: 就緒

位置:

註解:

☐ 列印到檔案(F)

頁面範圍

○ 全部(L)

○ 選取項目(T) ○ 目前頁面(U)

● 頁數(G): 1, 12-15

請輸入一個頁碼或單一分頁範圍。例如 5-12

份數(C): 1

☑ 自動分頁(O)