HIROSHIMA UNIVERSITY

Fundamental Data Science (30104001)

Lecture 5 — Visualizing data in R

Jorge N. Tendeiro
Hiroshima University

Today

How to use **R**:

- Import data.
- Compute some descriptive statistics (location, spread).
- Draw some plots (histograms, boxplots, scatterplots).

Today

How to use **R**:

- Import data.
- Compute some descriptive statistics (location, spread).
- Draw some plots (histograms, boxplots, scatterplots).

We will focus on using Posit Cloud.

Students can also, **optionally**, run R locally on their laptops.

Today

Before we start, make sure you:

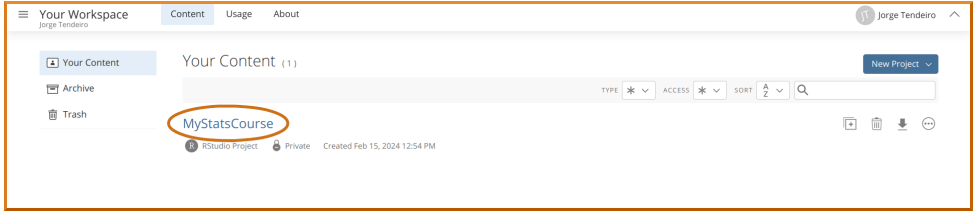
- Have created a folder called "Stat" on your desktop.
- Have downloaded the data file called `seiseki.csv` from *Moodle*, "Lecture 5".
- Have saved the data file in folder "Stat" on your desktop.
- Have registered your free account at RStudio Cloud (Posit Cloud).

Posit Cloud

Posit Cloud

I am expecting that you:

- Already saw the recommended video clip:
  https://vimeo.com/913207949
- Created a project for this course.
  (Mine is shown below; I called it "MyStatsCourse").

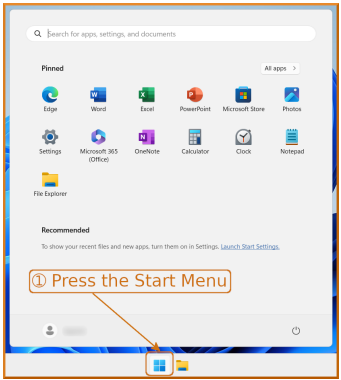


Now enter your project.

(In case you are running RStudio locally)

In case you installed R and RStudio on your computer, then simply start RStudio:

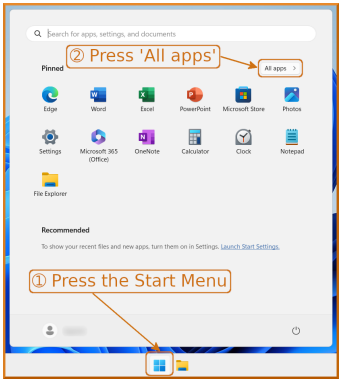
- Windows (Linux not much different):



(In case you are running RStudio locally)

In case you installed R and RStudio on your computer, then simply start RStudio:

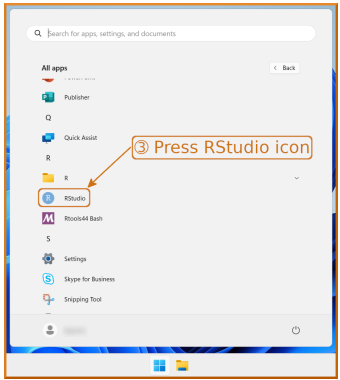
- Windows (Linux not much different):



(In case you are running RStudio locally)

In case you installed R and RStudio on your computer, then simply start RStudio:

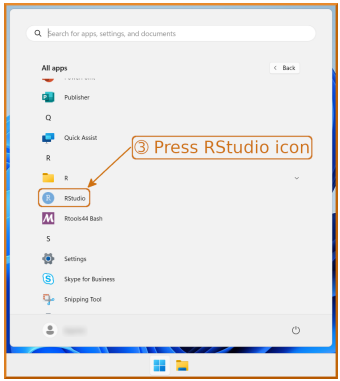
- Windows (Linux not much different):



(In case you are running RStudio locally)

In case you installed R and RStudio on your computer, then simply start RStudio:

- Windows (Linux not much different):

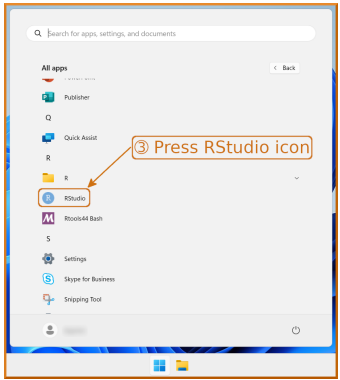


- Mac:

(In case you are running RStudio locally)

In case you installed R and RStudio on your computer, then simply start RStudio:

- Windows (Linux not much different):

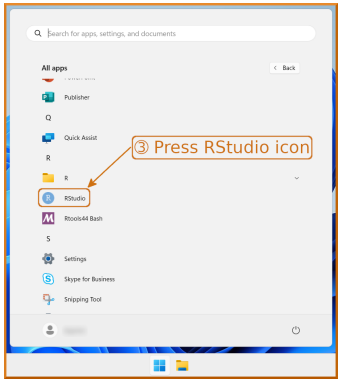


- Mac:

(In case you are running RStudio locally)

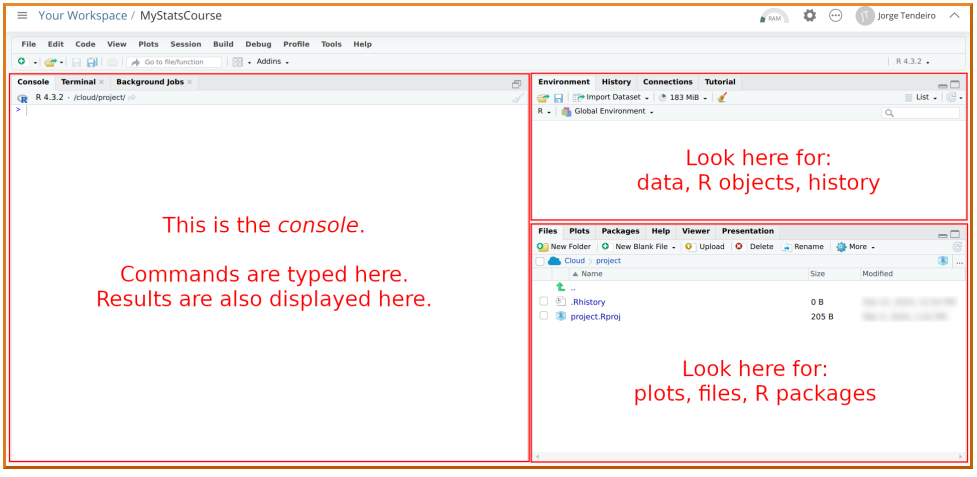
In case you installed R and RStudio on your computer, then simply start RStudio:

- Windows (Linux not much different):

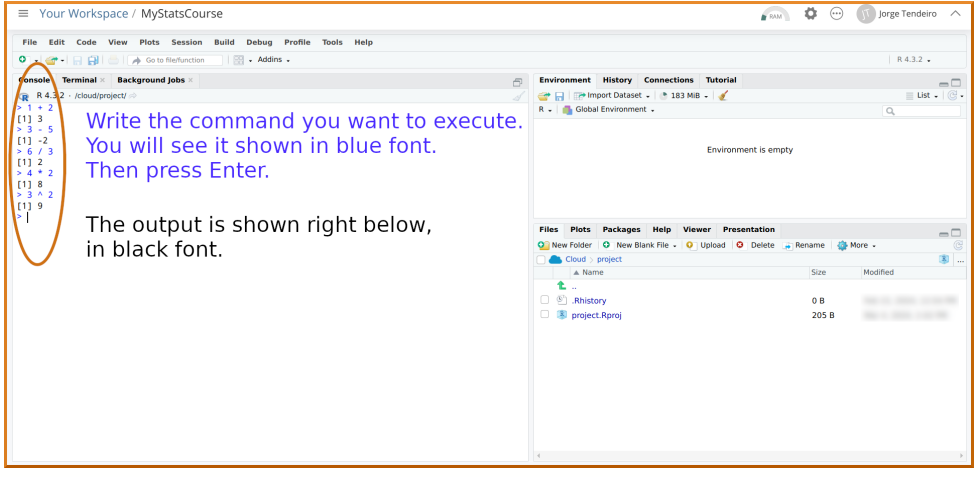


- Mac:

Posit Cloud — Project's main page

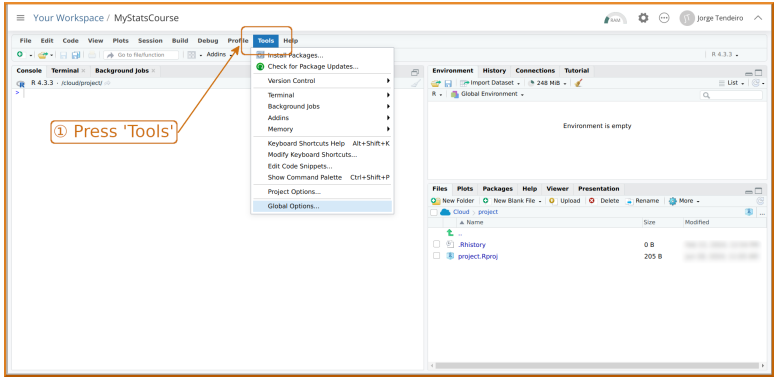


Posit Cloud — Basic calculations



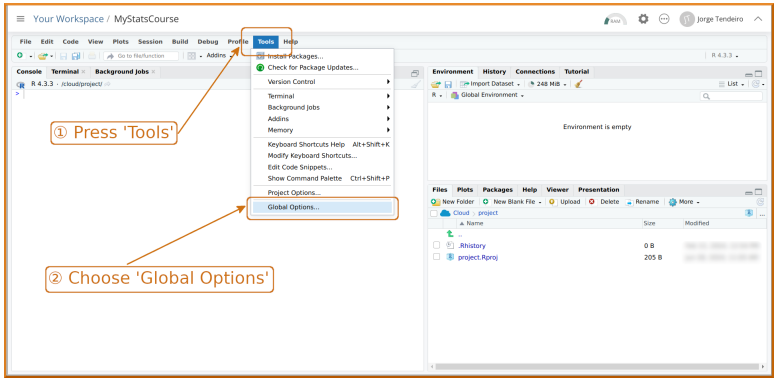
Posit Cloud — Adjusting font size

In case you need to adjust the screen's font size, do like this:



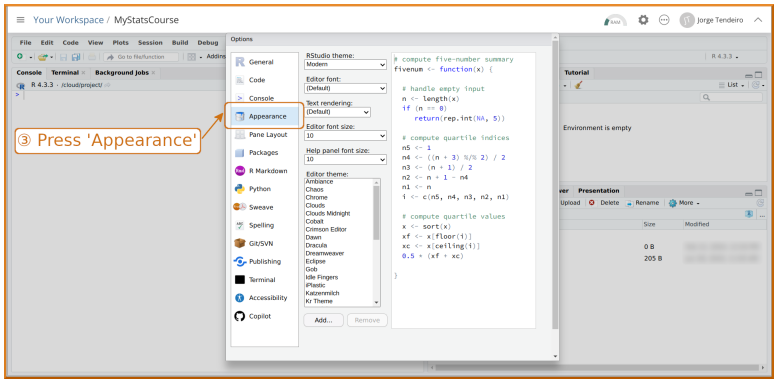
Posit Cloud — Adjusting font size

In case you need to adjust the screen's font size, do like this:



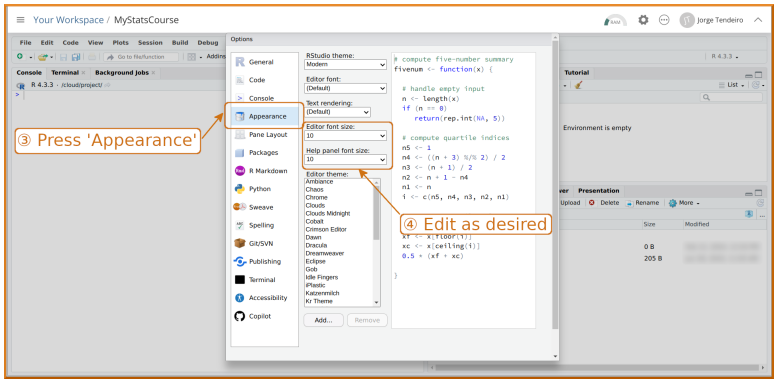
Posit Cloud — Adjusting font size

In case you need to adjust the screen's font size, do like this:



Posit Cloud — Adjusting font size

In case you need to adjust the screen's font size, do like this:



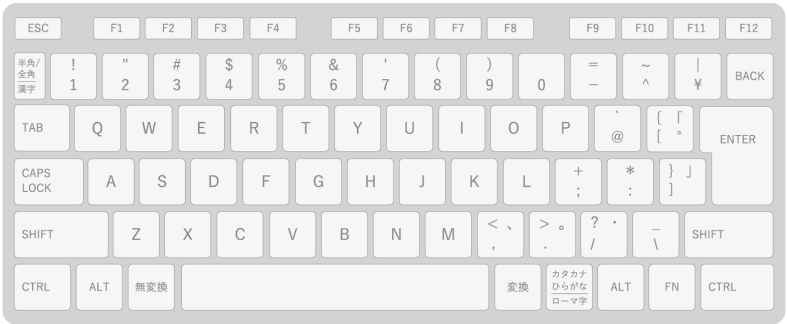
Posit Cloud — Typing

- RStudio is designed for half-width characters.
  You should make sure that the typing mode is set for half-width.

Posit Cloud — Typing

- RStudio is designed for half-width characters.
  You should make sure that the typing mode is set for half-width.
- Make sure you know how to type all required characters:

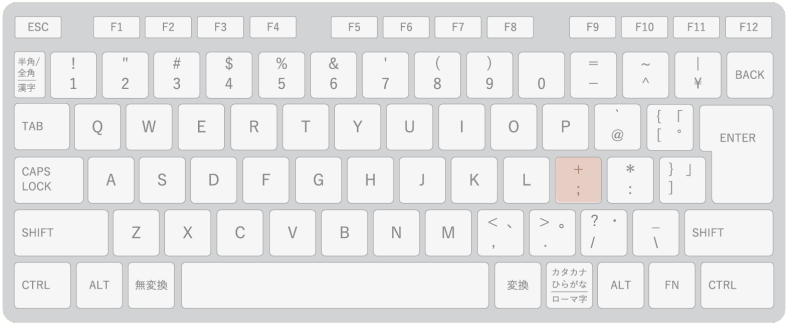
Keyboard.



Posit Cloud — Typing

- RStudio is designed for half-width characters.
  You should make sure that the typing mode is set for half-width.
- Make sure you know how to type all required characters:

The ; (semicolon) character.



Posit Cloud — Typing

- RStudio is designed for half-width characters.
  You should make sure that the typing mode is set for half-width.
- Make sure you know how to type all required characters:

The + (plus) character.

Posit Cloud — Typing

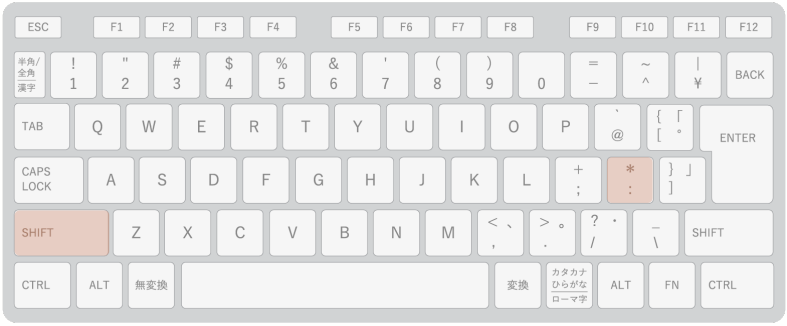
- RStudio is designed for half-width characters.
  You should make sure that the typing mode is set for half-width.
- Make sure you know how to type all required characters:

The : (colon) character.

Posit Cloud — Typing

- RStudio is designed for half-width characters.
  You should make sure that the typing mode is set for half-width.
- Make sure you know how to type all required characters:

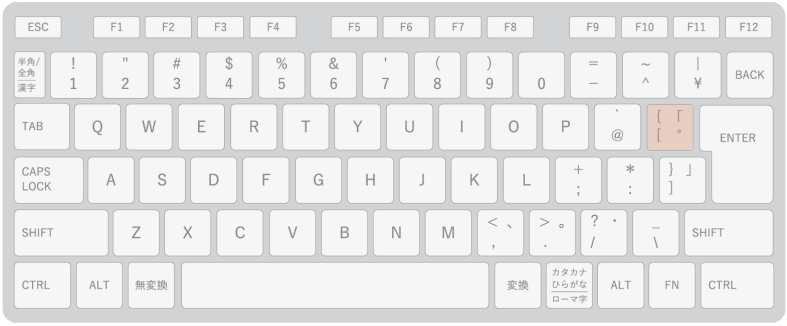
*The * (asterisk) character.*



Posit Cloud — Typing

- RStudio is designed for half-width characters.
  You should make sure that the typing mode is set for half-width.
- Make sure you know how to type all required characters:

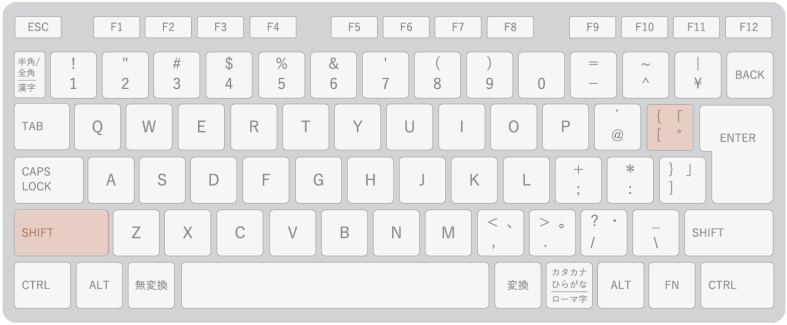
The [(open square brackets) character.



Posit Cloud — Typing

- RStudio is designed for half-width characters.
  You should make sure that the typing mode is set for half-width.
- Make sure you know how to type all required characters:

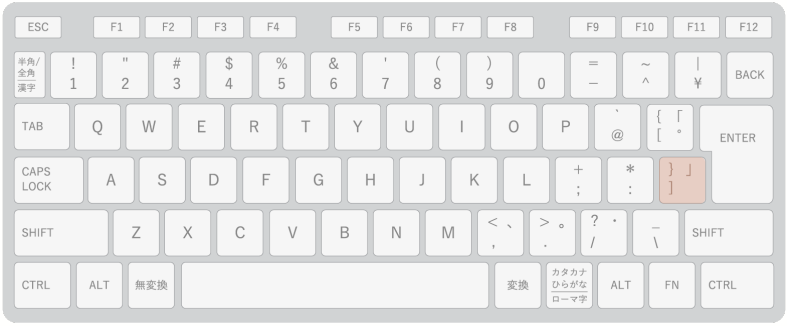
The { (open braces) character.



Posit Cloud — Typing

- RStudio is designed for half-width characters.
  You should make sure that the typing mode is set for half-width.
- Make sure you know how to type all required characters:

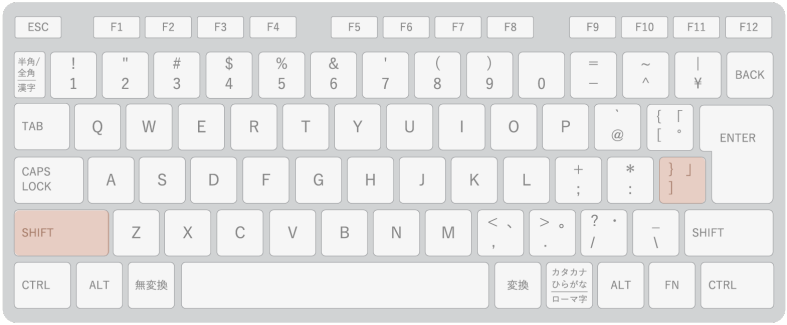
The] (close square brackets) character.



Posit Cloud — Typing

- RStudio is designed for half-width characters.
  You should make sure that the typing mode is set for half-width.
- Make sure you know how to type all required characters:

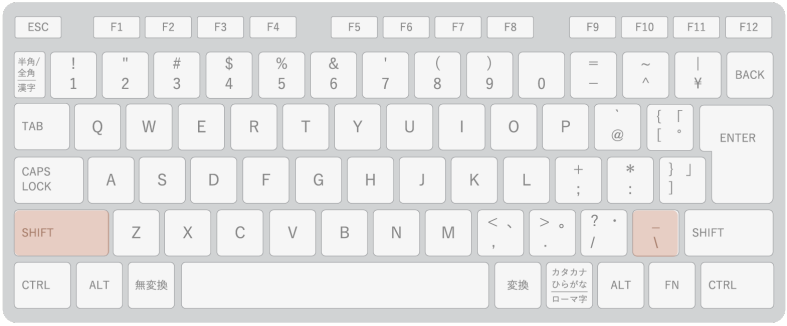
The } (close braces) character.



Posit Cloud — Typing

- RStudio is designed for half-width characters.
  You should make sure that the typing mode is set for half-width.
- Make sure you know how to type all required characters:

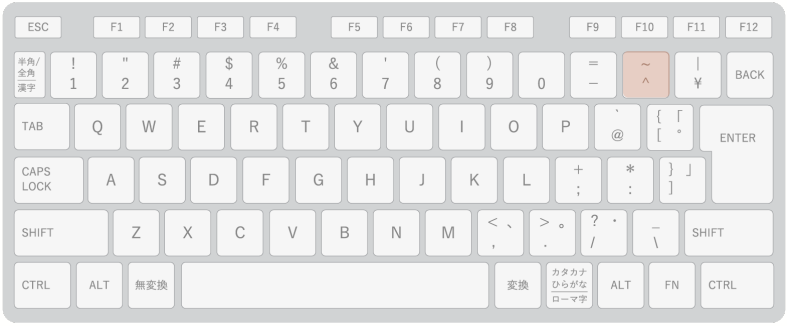
The _ (underscore) character.



Posit Cloud — Typing

- RStudio is designed for half-width characters.
  You should make sure that the typing mode is set for half-width.
- Make sure you know how to type all required characters:

The ^ (caret) character.



Posit Cloud — Typing

- RStudio is designed for half-width characters.
  You should make sure that the typing mode is set for half-width.
- Make sure you know how to type all required characters:

The ~ (tilde) character.

Posit Cloud — Typing

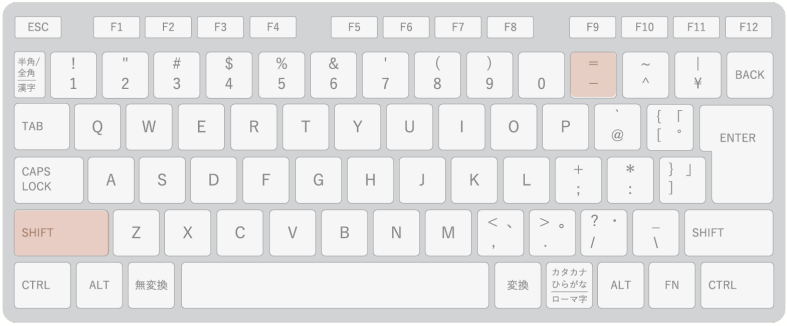
- RStudio is designed for half-width characters.
  You should make sure that the typing mode is set for half-width.
- Make sure you know how to type all required characters:

The - (minus) character.

Posit Cloud — Typing

- RStudio is designed for half-width characters.
  You should make sure that the typing mode is set for half-width.
- Make sure you know how to type all required characters:

The = (equal) character.



Posit Cloud — Typing

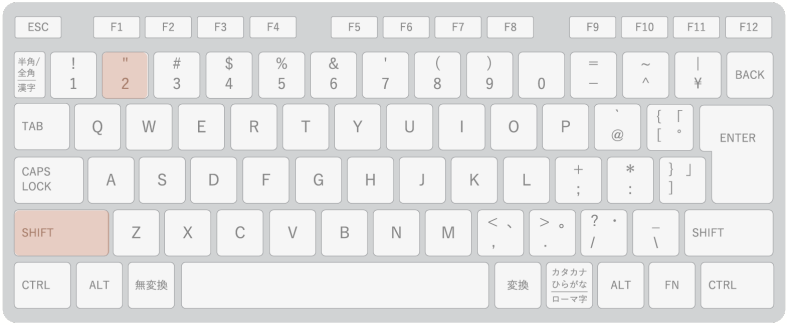
- RStudio is designed for half-width characters.
  You should make sure that the typing mode is set for half-width.
- Make sure you know how to type all required characters:

The ' (single quote) character.

Posit Cloud — Typing

- RStudio is designed for half-width characters.
  You should make sure that the typing mode is set for half-width.
- Make sure you know how to type all required characters:

The " (double quotes) character.



Posit Cloud — Using R scripts

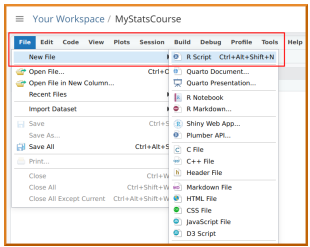
Commands executed from the console are **not** directly saved on a file.
This is fine for quick computations. But, when working on long assignments, it is much better to use files to save all the commands.
Such files are usually referred to as R scripts.

Posit Cloud — Using R scripts

Commands executed from the console are **not** directly saved on a file.
This is fine for quick computations. But, when working on long assignments, it is much better to use files to save all the commands.
Such files are usually referred to as R scripts.

You can create a new script in various, equivalent, ways:

- `File > New File > R Script`:

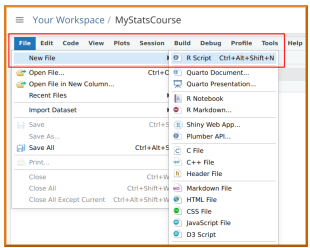


Posit Cloud — Using R scripts

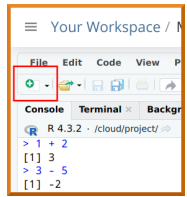
Commands executed from the console are **not** directly saved on a file.
This is fine for quick computations. But, when working on long assignments, it is much better to use files to save all the commands.
Such files are usually referred to as R scripts.

You can create a new script in various, equivalent, ways:

- `File > New File > R Script`:

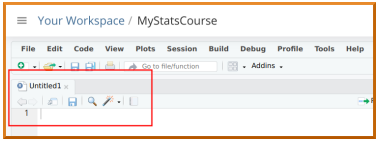


- Press the icon on the top-left.



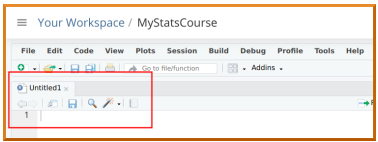
Posit Cloud — Using R scripts

Now the file is created:

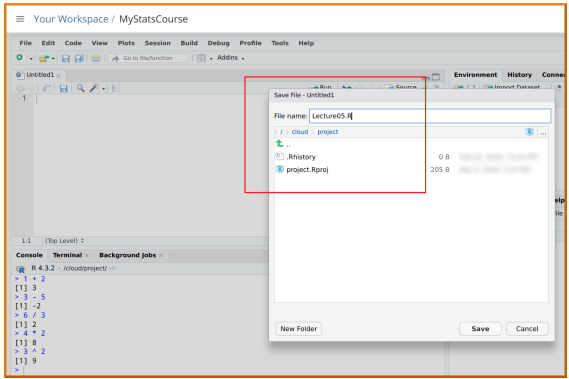


Posit Cloud — Using R scripts

Now the file is created:

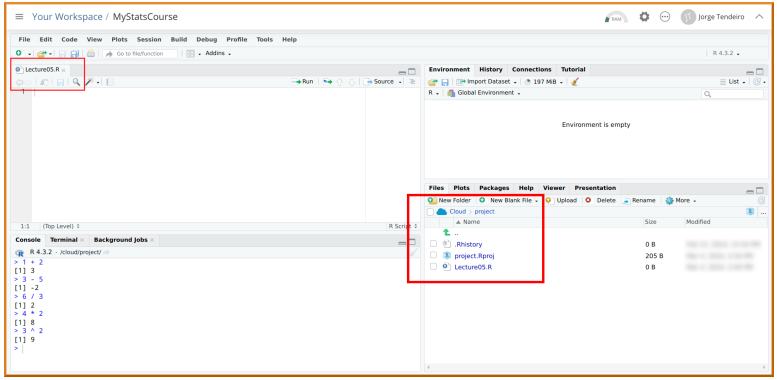


But it is **not yet saved**.
You can now save the file (`File > Save As`):



Posit Cloud — Using R scripts

Observe that the file is saved to your project's area at Posit Cloud.
The file is **not** on your computer!



We will see later how you can download files from Posit Cloud to your computer.

Exercise (1)

Calculate in R:

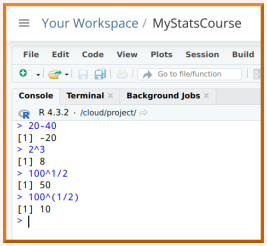
1. `20-40`.
2. 2 to the power of 3.
3. `100^1/2` and `100^(1/2)` and compare the results.

Exercise (1) — ANSWER

Calculate in R:

1. `20-40`.
2. 2 to the power of 3.
3. `100^1/2` and `100^(1/2)` and compare the results.

Answer

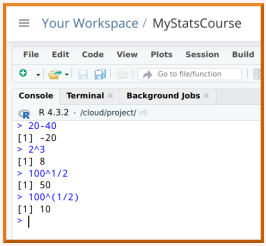


Exercise (1) — ANSWER

Calculate in R:

1. `20-40`.
2. 2 to the power of 3.
3. `100^1/2` and `100^(1/2)` and compare the results.

Answer



Note that

$$\texttt{100\^{}1/2} = (100^1)/2 = 50 \text{ and } \texttt{100\^{}(1/2)} = \sqrt{100} = 10.$$

It is very important to use parentheses in order to compute exactly what we want!

R input and output

Here on the slides I will use the following writing system to display R input and output code:

Input

```
20-40
```

Output

```
[1] -20
```

```
2^3
```

```
[1] 8
```

```
100^1/2
```

```
[1] 50
```

```
100^(1/2)
```

```
[1] 10
```

Note:
Use the discussion board in *Moodle* to ask questions whenever you get an error in R that you **cannot** understand.

Import data

Today's data

Today we will use the data file `seiseki.csv`:

Scores of 166 2nd year junior high students (Sugiyama et al., 2014).

You can get this file from *Moodle*, Lecture 5 folder.
Download it and put it in the "Stat" folder on your desktop.

Today's data

Today we will use the data file `seiseki.csv`:

Scores of 166 2nd year junior high students (Sugiyama et al., 2014).

You can get this file from *Moodle*, Lecture 5 folder.
Download it and put it in the "Stat" folder on your desktop.

	A	B	C	D	E	F	G	H	I	J
1	ID	kokugo	shakai	sugaku	rika	ongaku	bijutu	taiiku	gika	eigo
2	1	30	43	51	63	60	66	37	44	20
3	2	39	21	49	56	70	72	56	63	16
4	3	29	30	23	57	69	76	33	54	6
5	4	95	87	77	100	77	82	78	96	87
6	5	70	71	78	67	72	82	46	63	44
7	6	67	53	56	61	61	76	70	66	40
8	7	29	26	44	52	37	68	33	43	13
9	8	56	54	37	59	35	64	53	67	7
10	9	45	21	7	44	16	52	34	46	3
11	10	68	41	29	81	55	71	29	72	51
12	11	50	43	80	73	35	50	42	65	10
13	12	70	61	61	71	55	56	35	67	33

Variable	Description
ID	ID number
kokugo	Japanese
shakai	Society
sugaku	Math

Variable	Description
rika	Science
ongaku	Music
bijutu	Art
taiiku	PE

Variable	Description
gika	Industrial arts and homemaking
eigo	English

Important info about data files

- Use half-width alphanumeric characters for all variable names.

	A	B	C	D	E	F	G	H	I	J
1	ID	kokugo	shakai	sugaku	rika	ongaku	bijutu	taiiku	gika	eigo
2	1	30	43	51	63	60	66	37	44	20
3	2	39	21	49	56	70	72	56	63	16
4	3	29	30	23	57	69	76	33	54	6
5	4	95	87	77	100	77	82	78	96	87
6	5	70	71	78	67	72	82	46	63	44
7	6	67	53	56	61	61	76	70	66	40
8	7	29	26	44	52	37	68	33	43	13
9	8	56	54	37	59	35	64	53	67	7
10	9	45	21	7	44	16	52	34	46	3
11	10	68	41	29	81	55	71	29	72	51
12	11	50	43	80	73	35	50	42	65	10
13	12	70	61	61	71	55	56	35	67	33

Important info about data files

- Use half-width alphanumeric characters for all variable names.
- Give a name to each variable.
  All variables names should be place in the top row.

	A	B	C	D	E	F	G	H	I	J
1	ID	kokugo	shakai	sugaku	rika	ongaku	bijutu	taiiku	gika	eigo
2	1	30	43	51	63	60	66	37	44	20
3	2	39	21	49	56	70	72	56	63	16
4	3	29	30	23	57	69	76	33	54	6
5	4	95	87	77	100	77	82	78	96	87
6	5	70	71	78	67	72	82	46	63	44
7	6	67	53	56	61	61	76	70	66	40
8	7	29	26	44	52	37	68	33	43	13
9	8	56	54	37	59	35	64	53	67	7
10	9	45	21	7	44	16	52	34	46	3
11	10	68	41	29	81	55	71	29	72	51
12	11	50	43	80	73	35	50	42	65	10
13	12	70	61	61	71	55	56	35	67	33

Important info about data files

- Use half-width alphanumeric characters for all variable names.
- Give a name to each variable.
  All variables names should be place in the top row.
- Data values start from the second row onward.

	A	B	C	D	E	F	G	H	I	J
1	ID	kokugo	shakai	sugaku	rika	ongaku	bijutu	taiiku	gika	eigo
2	1	30	43	51	63	60	66	37	44	20
3	2	39	21	49	56	70	72	56	63	16
4	3	29	30	23	57	69	76	33	54	6
5	4	95	87	77	100	77	82	78	96	87
6	5	70	71	78	67	72	82	46	63	44
7	6	67	53	56	61	61	76	70	66	40
8	7	29	26	44	52	37	68	33	43	13
9	8	56	54	37	59	35	64	53	67	7
10	9	45	21	7	44	16	52	34	46	3
11	10	68	41	29	81	55	71	29	72	51
12	11	50	43	80	73	35	50	42	65	10

Important info about data files

- Use half-width alphanumeric characters for all variable names.
- Give a name to each variable.
  All variables names should be place in the top row.
- Data values start from the second row onward.
- Make sure to **delete** all other irrelevant rows!
  This is to prevent errors when importing data into R.

	A	B	C	D	E	F	G	H	I	J
1	ID	kokugo	shakai	sugaku	rika	ongaku	bijutu	taiiku	gika	eigo
2	1	30	43	51	63	60	66	37	44	20
3	2	39	21	49	56	70	72	56	63	16
4	3	29	30	23	57	69	76	33	54	6
5	4	95	87	77	100	77	82	78	96	87
6	5	70	71	78	67	72	82	46	63	44
7	6	67	53	56	61	61	76	70	66	40
8	7	29	26	44	52	37	68	33	43	13
9	8	56	54	37	59	35	64	53	67	7
10	9	45	21	7	44	16	52	34	46	3
11	10	68	41	29	81	55	71	29	72	51
12	11	50	43	80	73	35	50	42	65	10

Important info about data files

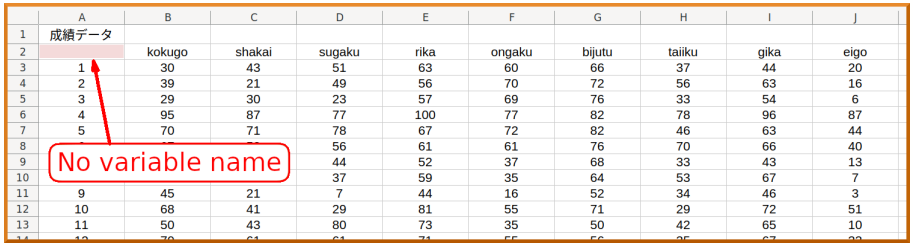
- Use half-width alphanumeric characters for all variable names.
- Give a name to each variable.
  All variables names should be place in the top row.
- Data values start from the second row onward.
- Make sure to **delete** all other irrelevant rows!
  This is to prevent errors when importing data into R.
- Even if you manage to import data from a file, you will typically still need to fix some issues later on.

	A	B	C	D	E	F	G	H	I	J
1	ID	kokugo	shakai	sugaku	rika	ongaku	bijutu	taiiku	gika	eigo
2	1	30	43	51	63	60	66	37	44	20
3	2	39	21	49	56	70	72	56	63	16
4	3	29	30	23	57	69	76	33	54	6
5	4	95	87	77	100	77	82	78	96	87
6	5	70	71	78	67	72	82	46	63	44
7	6	67	53	56	61	61	76	70	66	40
8	7	29	26	44	52	37	68	33	43	13
9	8	56	54	37	59	35	64	53	67	7
10	9	45	21	7	44	16	52	34	46	3
11	10	68	41	29	81	55	71	29	72	51
12	11	50	43	80	73	35	50	42	65	10

Data file — Bad example

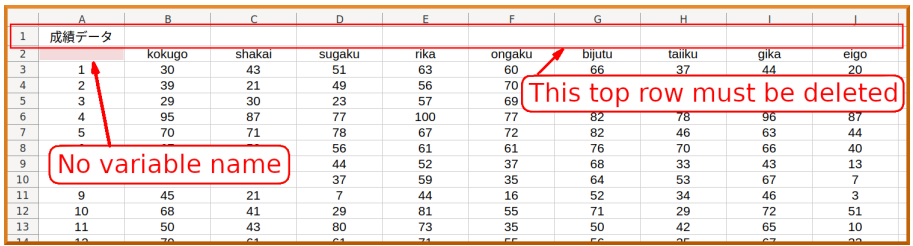
	A	B	C	D	E	F	G	H	I	J
1	成績データ									
2		kokugo	shakai	sugaku	rika	ongaku	bijutu	taiiku	gika	eigo
3	1	30	43	51	63	60	66	37	44	20
4	2	39	21	49	56	70	72	56	63	16
5	3	29	30	23	57	69	76	33	54	6
6	4	95	87	77	100	77	82	78	96	87
7	5	70	71	78	67	72	82	46	63	44
8	6	67	53	56	61	61	76	70	66	40
9	7	29	26	44	52	37	68	33	43	13
10	8	56	54	37	59	35	64	53	67	7
11	9	45	21	7	44	16	52	34	46	3
12	10	68	41	29	81	55	71	29	72	51
13	11	50	43	80	73	35	50	42	65	10

Data file — Bad example



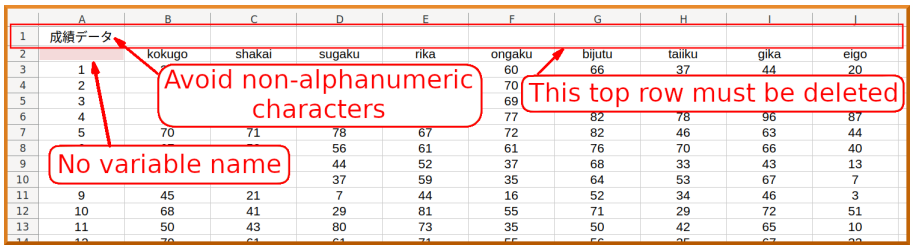
	A	B	C	D	E	F	G	H	I	J
1	成績データ									
2		kokugo	shakai	sugaku	rika	ongaku	bijutu	taiiku	gika	eigo
3	1	30	43	51	63	60	66	37	44	20
4	2	39	21	49	56	70	72	56	63	16
5	3	29	30	23	57	69	76	33	54	6
6	4	95	87	77	100	77	82	78	96	87
7	5	70	71	78	67	72	82	46	63	44
8				56	61	61	76	70	66	40
9				44	52	37	68	33	43	13
10				37	59	35	64	53	67	7
11	9	45	21	7	44	16	52	34	46	3
12	10	68	41	29	81	55	71	29	72	51
13	11	50	43	80	73	35	50	42	65	10

Data file — Bad example



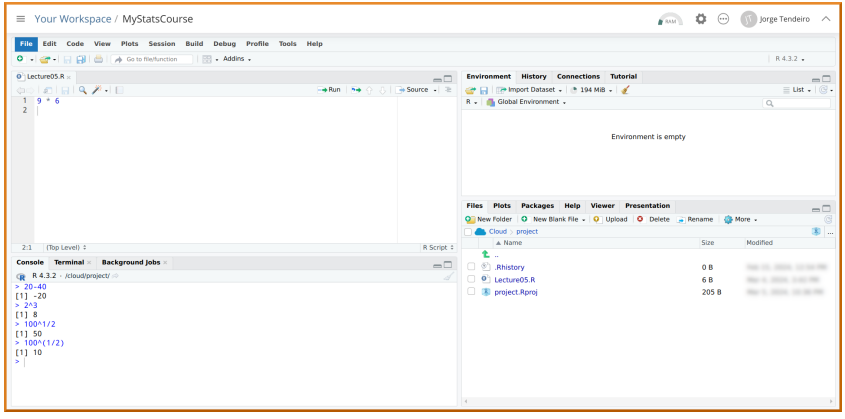
	A	B	C	D	E	F	G	H	I	J
1	成績データ									
2		kokugo	shakai	sugaku	rika	ongaku	bijutu	taiiku	gika	eigo
3	1	30	43	51	63	60	66	37	44	20
4	2	39	21	49	56	70				
5	3	29	30	23	57	69				
6	4	95	87	77	100	77	82	78	96	87
7	5	70	71	78	67	72	82	46	63	44
8				56	61	61	76	70	66	40
9				44	52	37	68	33	43	13
10				37	59	35	64	53	67	7
11	9	45	21	7	44	16	52	34	46	3
12	10	68	41	29	81	55	71	29	72	51
13	11	50	43	80	73	35	50	42	65	10

Data file — Bad example

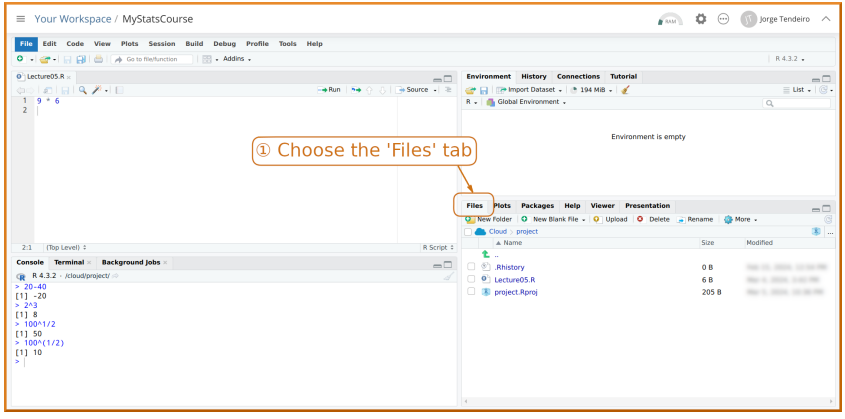


	A	B	C	D	E	F	G	H	I	J
1	成績データ									
2		kokugo	shakai	sugaku	rika	ongaku	bijutu	taiiku	gika	eigo
3	1					60	66	37	44	20
4	2					70				
5	3					69				
6	4					77	82	78	96	87
7	5	70	71	78	67	72	82	46	63	44
8				56	61	61	76	70	66	40
9				44	52	37	68	33	43	13
10				37	59	35	64	53	67	7
11	9	45	21	7	44	16	52	34	46	3
12	10	68	41	29	81	55	71	29	72	51
13	11	50	43	80	73	35	50	42	65	10

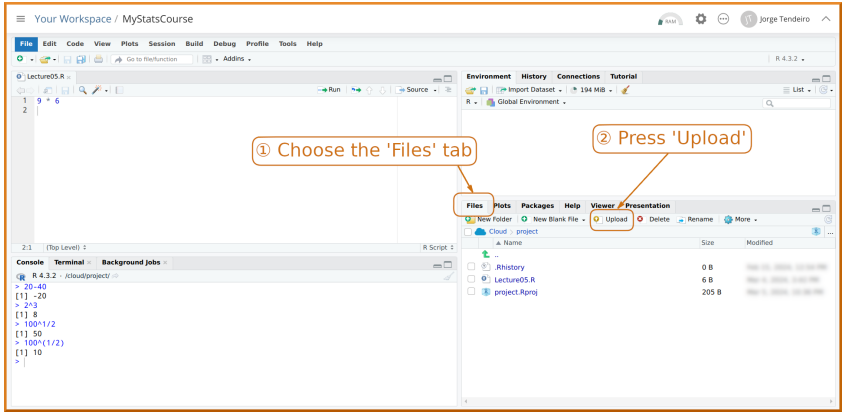
Uploading the data file to Posit Cloud



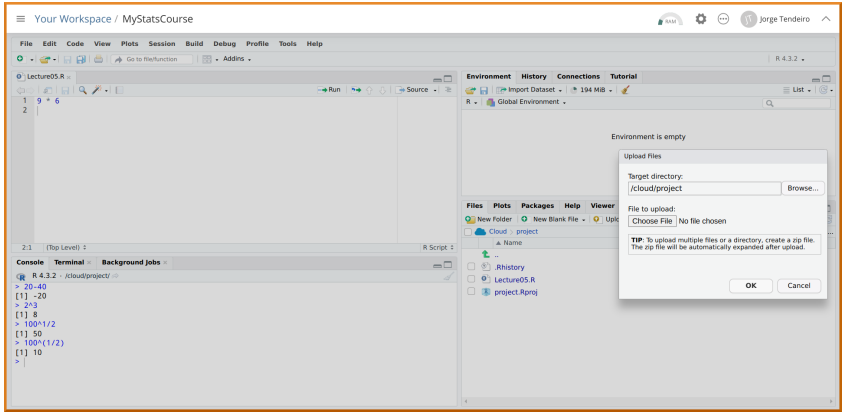
Uploading the data file to Posit Cloud



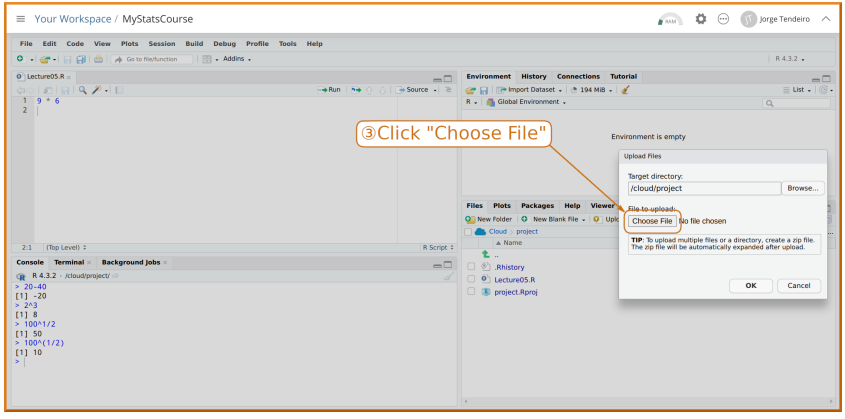
Uploading the data file to Posit Cloud



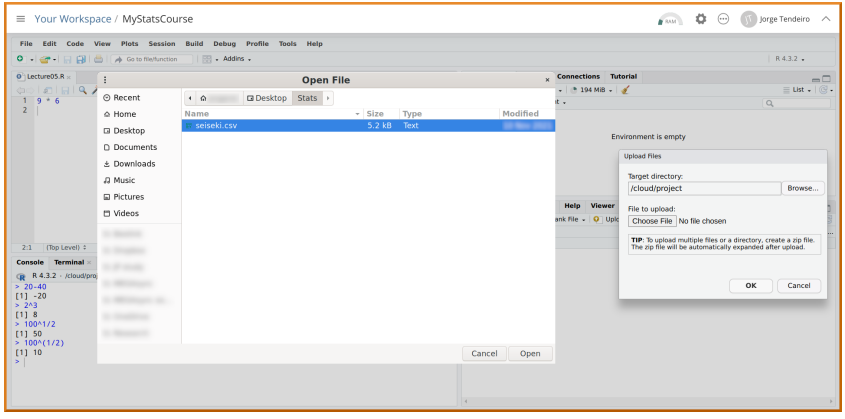
Uploading the data file to Posit Cloud



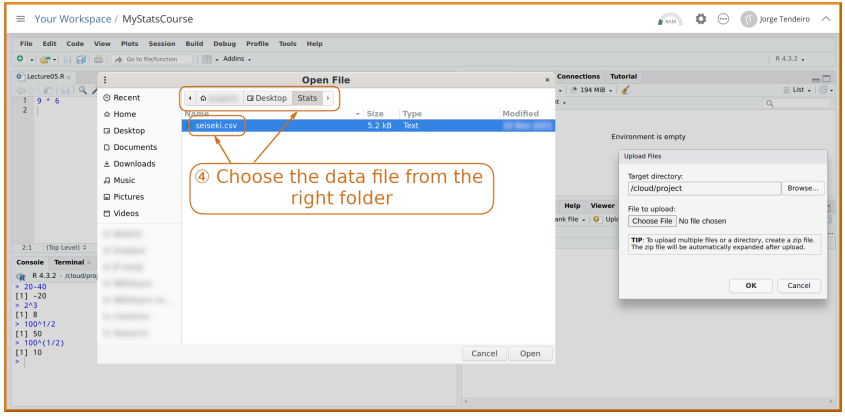
Uploading the data file to Posit Cloud



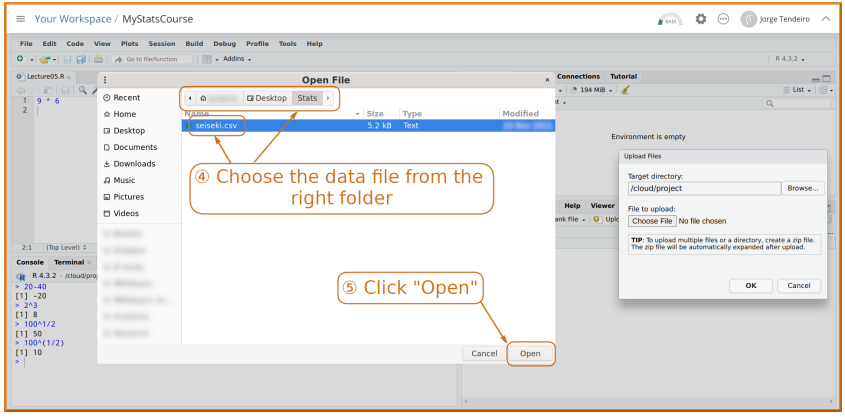
Uploading the data file to Posit Cloud



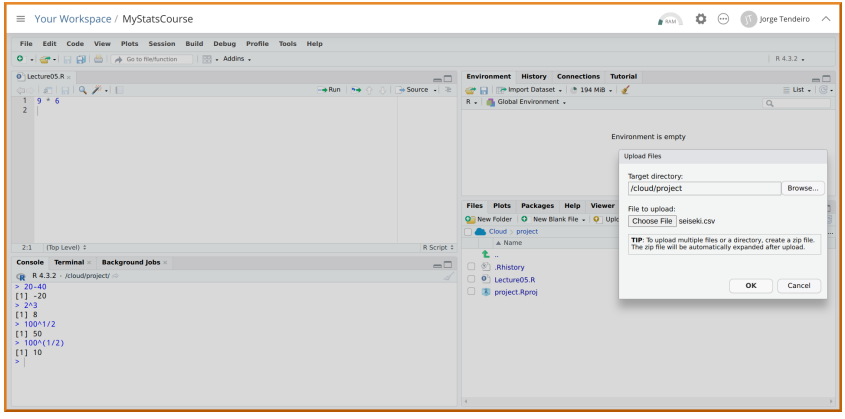
Uploading the data file to Posit Cloud



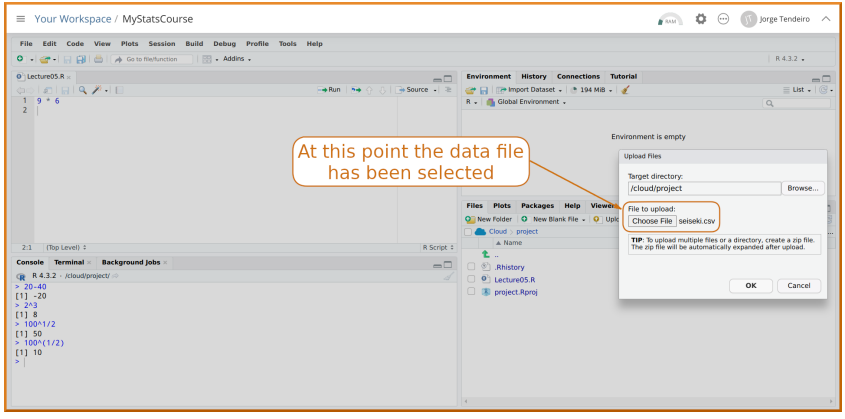
Uploading the data file to Posit Cloud



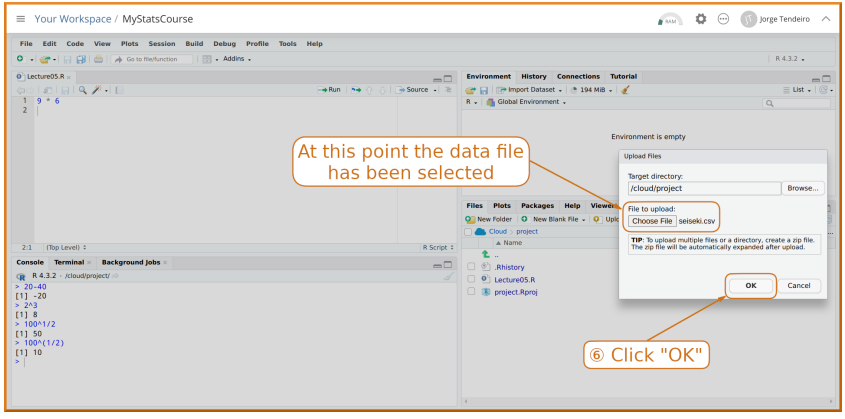
Uploading the data file to Posit Cloud



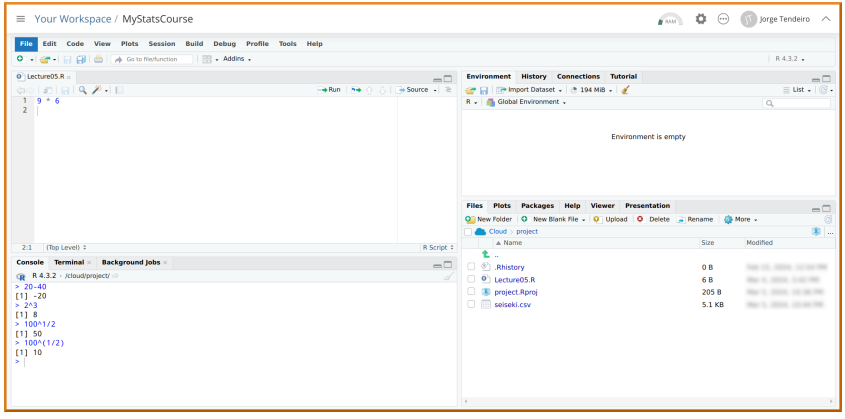
Uploading the data file to Posit Cloud



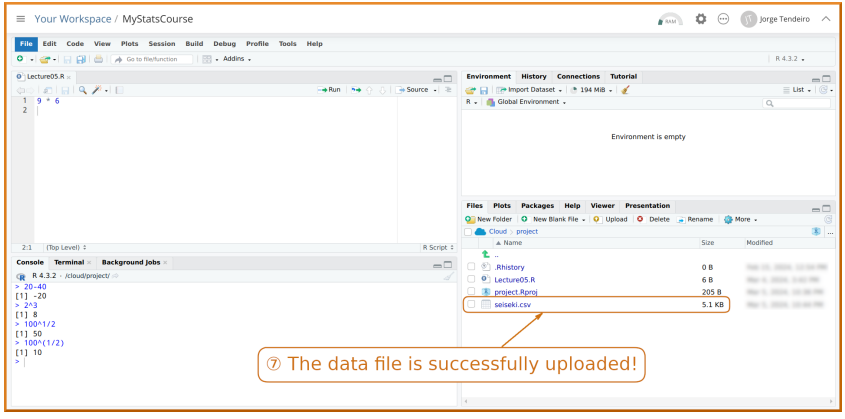
Uploading the data file to Posit Cloud



Uploading the data file to Posit Cloud



Uploading the data file to Posit Cloud



A data file needs only to be uploaded once.
Even if you log out and then log back in to Posit Cloud, the data file is still there.

For students running RStudio on their computers (not Posit Cloud)

Running RStudio locally

It is best to set a working directory.
In this way, RStudio knows exactly which folder to use to read/write data files.

Running RStudio locally

It is best to set a working directory.
In this way, RStudio knows exactly which folder to use to read/write data files.

In our case, the folder we want to use is `~/Desktop/Stats`.
So let's set this folder as the working directory.

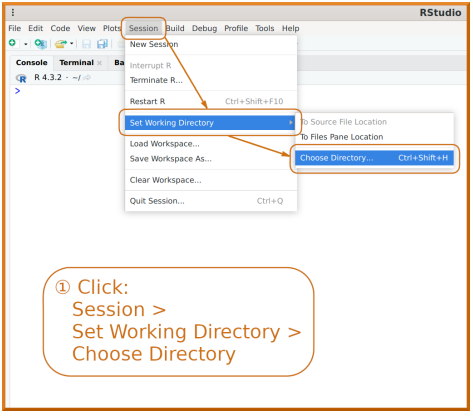
Running RStudio locally

It is best to set a working directory.
In this way, RStudio knows exactly which folder to use to read/write data files.

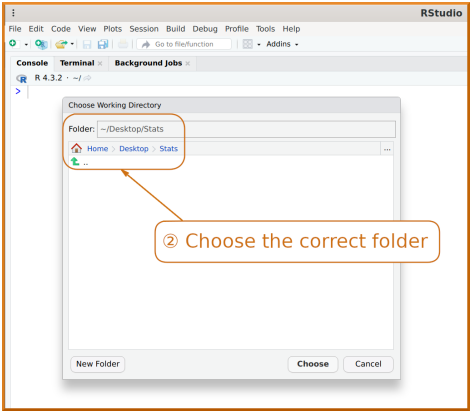
In our case, the folder we want to use is `~/Desktop/Stats`.
So let's set this folder as the working directory.

We'll go through two ways to do this.
You either use the first **or** the second procedure (no need for both!).

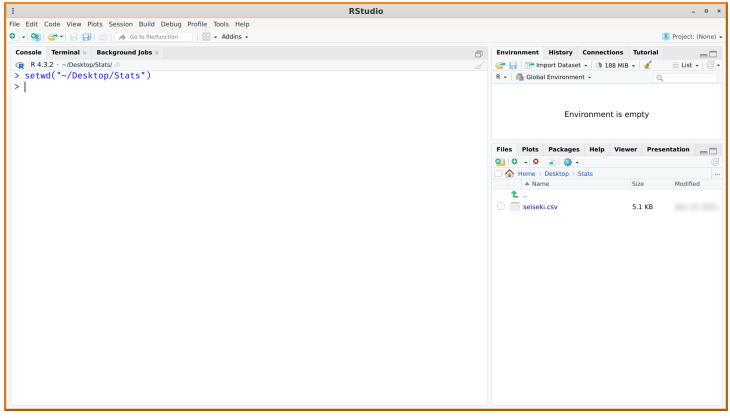
1 — Set up the working directory manually



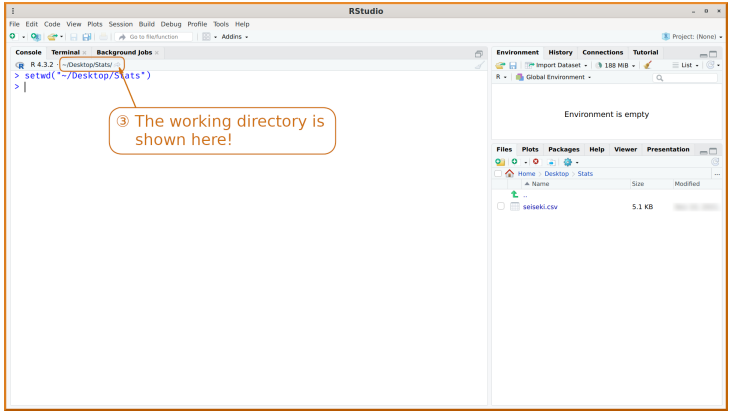
1 — Set up the working directory manually



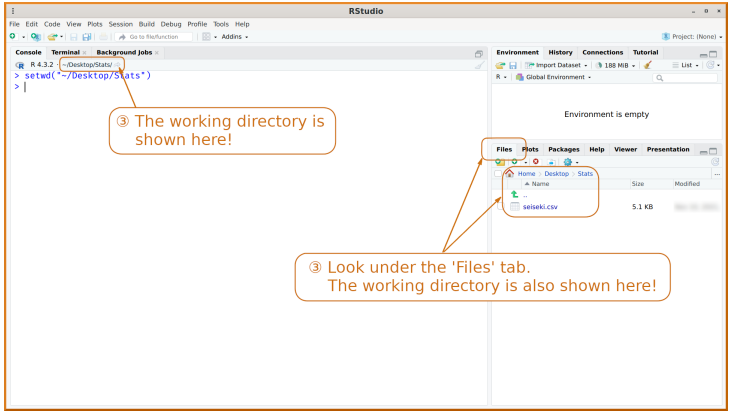
1 — Set up the working directory manually



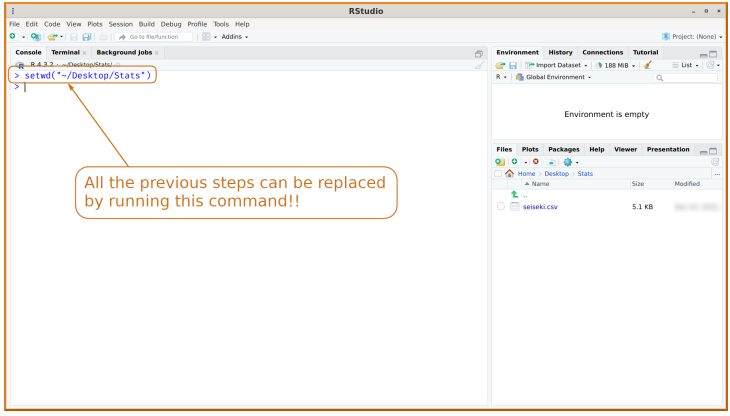
1 — Set up the working directory manually



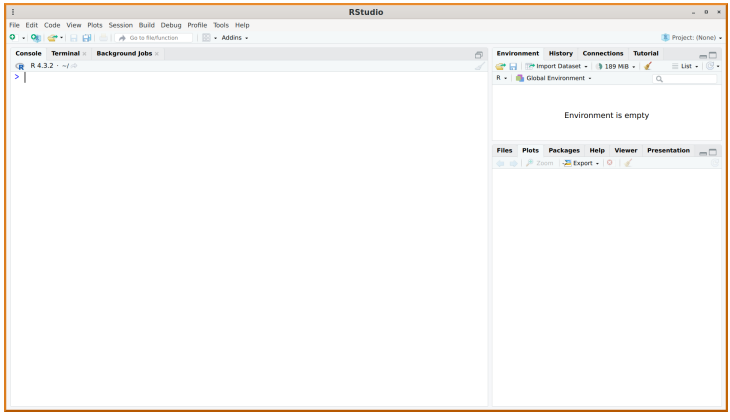
1 — Set up the working directory manually



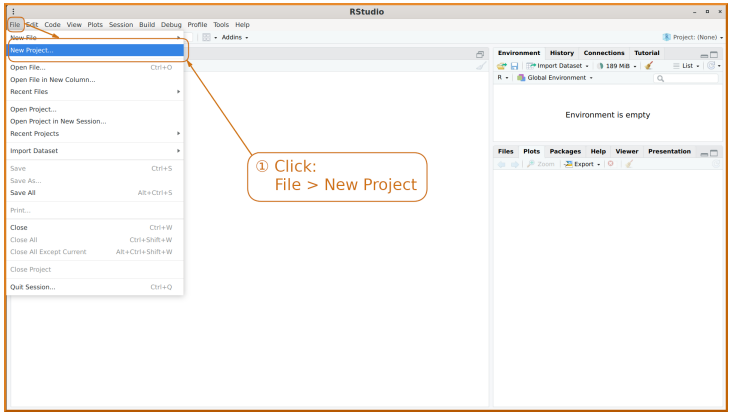
1 — Set up the working directory manually



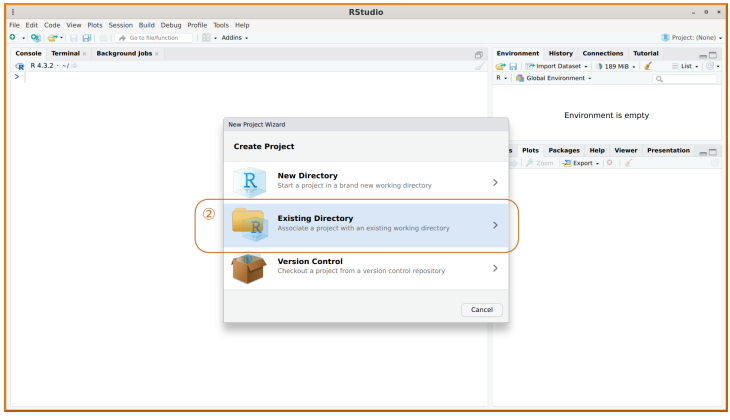
2 — Create a project



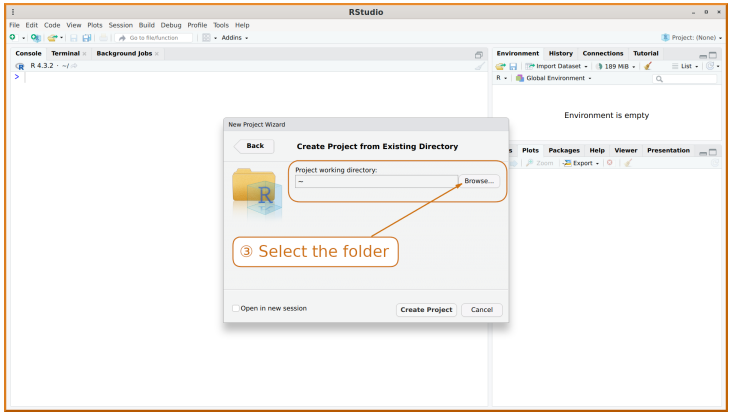
2 — Create a project



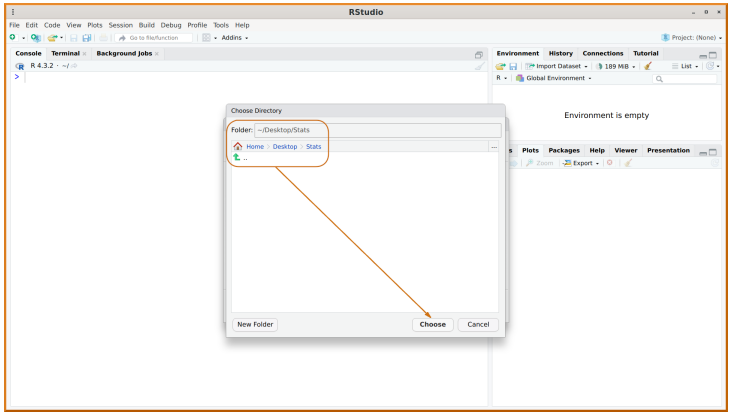
2 — Create a project



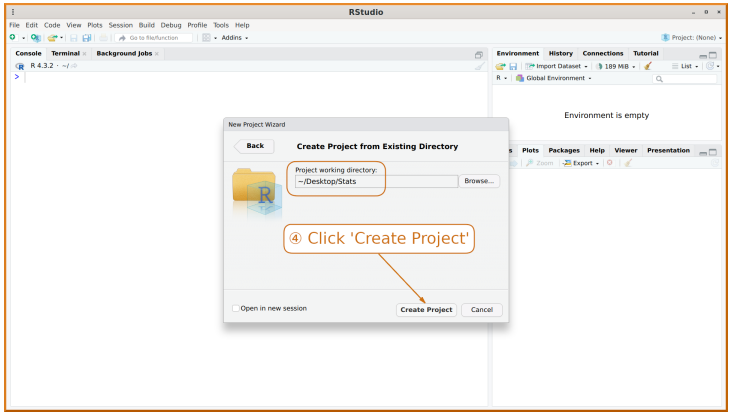
2 — Create a project



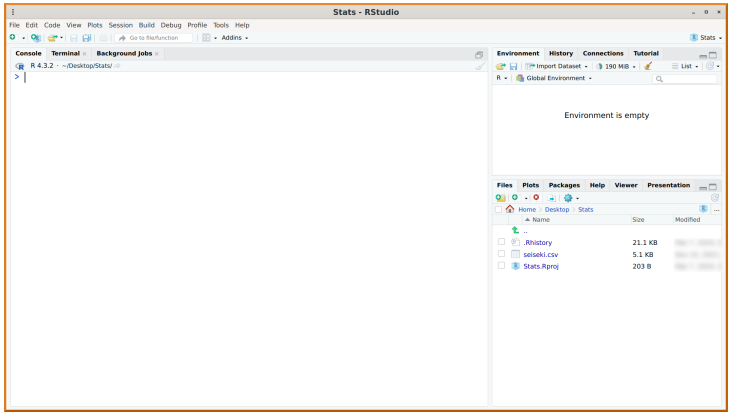
2 — Create a project



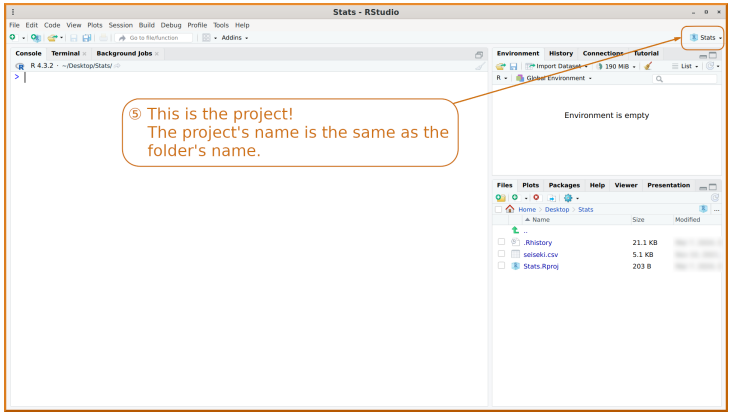
2 — Create a project



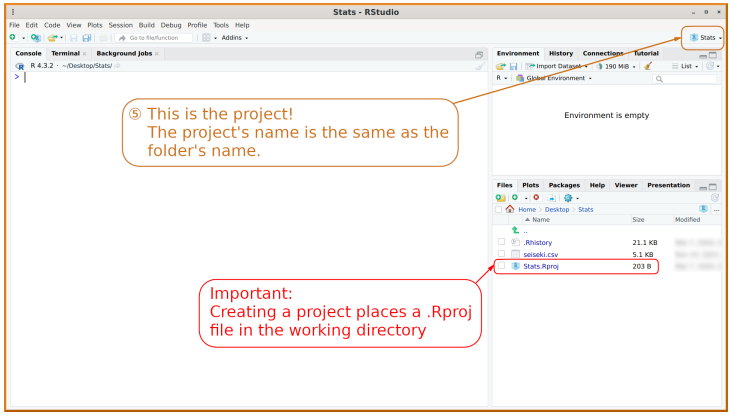
2 — Create a project



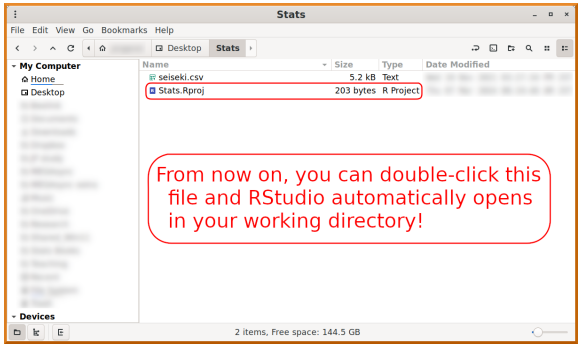
2 — Create a project



2 — Create a project



2 — Create a project



Load data in RStudio

Compute descriptive statistics (location, spread)

Load data in RStudio

The data file, `seiseki.csv`, is now available in your working directory.
However, the data are not yet loaded.

To load the data into your R session run the following command:

Input

```
seiseki ← read.csv("seiseki.csv")
```

Load data in RStudio

The data file, `seiseki.csv`, is now available in your working directory.
However, the data are not yet loaded.

To load the data into your R session run the following command:

Input

```
seiseki ← read.csv("seiseki.csv")
```

No output is displayed.
But the data set is now visible as an R object in the environment (top-right).

Load data in RStudio

The data file, `seiseki.csv`, is now available in your working directory.
However, the data are not yet loaded.

To load the data into your R session run the following command:

Input

```
seiseki ← read.csv("seiseki.csv")
```

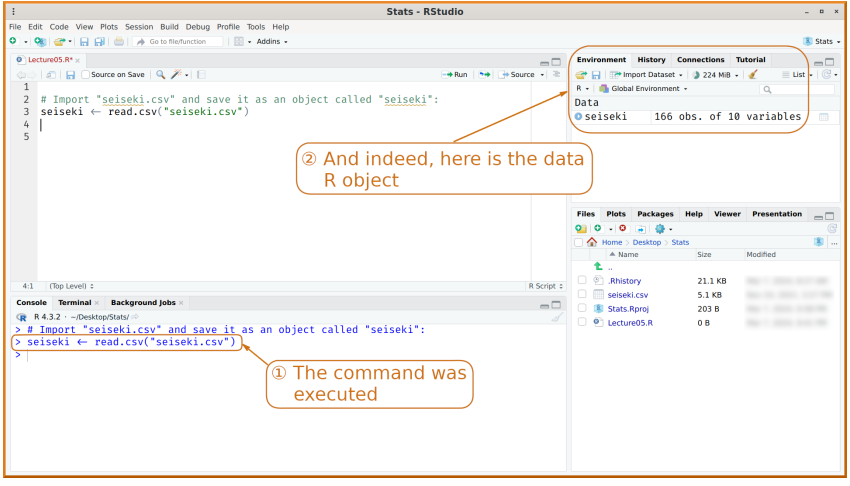
No output is displayed.
But the data set is now visible as an R object in the environment (top-right).

Tip: For importing text files with extension `.txt` try executing the following command:

Input

```
mydata ← read.table("FILENAME.txt") # the name 'mydata' can be anything you want
```

Load data in RStudio



Look at the data

Input

```
# Look at the top rows of 'seiseki':
head(seiseki)
```

Look at the data

Input

```
# Look at the top rows of 'seiseki':
head(seiseki)
```

Output

```
  ID kokugo shakai sugaku rika ongaku bijutu taiiku gika eigo
1  1     30     43     51   63     60     66     37   44   20
2  2     39     21     49   56     70     72     56   63   16
3  3     29     30     23   57     69     76     33   54    6
4  4     95     87     77  100     77     82     78   96   87
5  5     70     71     78   67     72     82     46   63   44
6  6     67     53     56   61     61     76     70   66   40
```

Look at the data

Input

```
# Look at the top rows of 'seiseki':
head(seiseki)
```

Output

```
  ID kokugo shakai sugaku rika ongaku bijutu taiiku gika eigo
1  1     30     43     51   63     60     66     37   44   20
2  2     39     21     49   56     70     72     56   63   16
3  3     29     30     23   57     69     76     33   54    6
4  4     95     87     77  100     77     82     78   96   87
5  5     70     71     78   67     72     82     46   63   44
6  6     67     53     56   61     61     76     70   66   40
```

Input

```
# What are the variable names?
colnames(seiseki)
```

Look at the data

Input

```
# Look at the top rows of 'seiseki':
head(seiseki)
```

Output

```
  ID kokugo shakai sugaku rika ongaku bijutu taiiku gika eigo
1  1     30     43     51   63     60     66     37   44   20
2  2     39     21     49   56     70     72     56   63   16
3  3     29     30     23   57     69     76     33   54    6
4  4     95     87     77  100     77     82     78   96   87
5  5     70     71     78   67     72     82     46   63   44
6  6     67     53     56   61     61     76     70   66   40
```

Input

```
# What are the variable names?
colnames(seiseki)
```

Output

```
[1] "ID"     "kokugo" "shakai" "sugaku"
[5] "rika"   "ongaku" "bijutu" "taiiku"
[9] "gika"   "eigo"
```

Look at the data

Input

```
# Look at the top rows of 'seiseki':
head(seiseki)
```

Output

```
  ID kokugo shakai sugaku rika ongaku bijutu taiiku gika eigo
1  1     30     43     51   63     60     66     37   44   20
2  2     39     21     49   56     70     72     56   63   16
3  3     29     30     23   57     69     76     33   54    6
4  4     95     87     77  100     77     82     78   96   87
5  5     70     71     78   67     72     82     46   63   44
6  6     67     53     56   61     61     76     70   66   40
```

Input

```
# What are the variable names?
colnames(seiseki)
```

Output

```
[1] "ID"     "kokugo" "shakai" "sugaku"
[5] "rika"   "ongaku" "bijutu" "taiiku"
[9] "gika"   "eigo"
```

Note: In R, all code written after # (the hash symbol) is not considered a command.

Extract data

The R object `seiseki` is a data frame:

Input

```
class(seiseki)
```

Extract data

The R object `seiseki` is a data frame:

Input

```
class(seiseki)
```

Output

```
[1] "data.frame"
```

Extract data

The R object `seiseki` is a data frame:

Input

```
class(seiseki)
```

Output

```
[1] "data.frame"
```

There are multiple ways to extract variables (=columns) from a data frame.
Here are a few options. All achieve the same: Extracting variable "`kokugo`":

Input

```
# See all variable names (and positions):
colnames(seiseki)

# Option 1 - Using the dollar notation:
seiseki$kokugo

# Option 2 - By variable name
#  (nothing before the comma = ALL rows):
seiseki[, "kokugo"]

# Option 3 - By position
#  ('kokugo' is the 2nd variable):
seiseki[, 2]
```

Extract data

The R object `seiseki` is a data frame:

Input

```
class(seiseki)
```

Output

```
[1] "data.frame"
```

There are multiple ways to extract variables (=columns) from a data frame.
Here are a few options. All achieve the same: Extracting variable "kokugo":

Input

```
# See all variable names (and positions):
colnames(seiseki)

# Option 1 - Using the dollar notation:
seiseki$kokugo

# Option 2 - By variable name
#  (nothing before the comma = ALL rows):
seiseki[, "kokugo"]

# Option 3 - By position
#  ('kokugo' is the 2nd variable):
seiseki[, 2]
```

Output

```
 [1] 30  39  29  95  70  67  29  56  45
[10] 68  50  70  46  23  77  15  37  55
[19] 0   35  60  70  82  44  54  80  46
[28] 48  51  37  31  35  50  33  27  36
[37] 77  61  84  78  7   57  60  17  82
[46] 37  40  72  74  87  66  90  43  24
[55] 41  42  33  29  83  73  66  71  62
[64] 47  53  61  21  74  72  50  5   32
[73] 23  34  70  9   61  34  38  72  ...
```

Some descriptives statistics

Input

```
# Mean:
mean(seiseki$kokugo)
```

Some descriptives statistics

Input

```
# Mean:
mean(seiseki$kokugo)
```

Output

```
[1] 52.33133
```

Some descriptives statistics

Input

```
# Mean:
mean(seiseki$kokugo)
```

Output

```
[1] 52.33133
```

Input

```
# Median:
median(seiseki$kokugo)
```

Some descriptives statistics

Input

```
# Mean:
mean(seiseki$kokugo)
```

Output

```
[1] 52.33133
```

Input

```
# Median:
median(seiseki$kokugo)
```

Output

```
[1] 53
```

Some descriptives statistics

Input

```
# Mean:
mean(seiseki$kokugo)
```

Output

```
[1] 52.33133
```

Input

```
# Median:
median(seiseki$kokugo)
```

Output

```
[1] 53
```

Input

```
# Minimum, quartiles (Hinge method), maximum:
fivenum(seiseki$kokugo)
```

Some descriptives statistics

Input

```
# Mean:
mean(seiseki$kokugo)
```

Output

```
[1] 52.33133
```

Input

```
# Median:
median(seiseki$kokugo)
```

Output

```
[1] 53
```

Input

```
# Minimum, quartiles (Hinge method), maximum:
fivenum(seiseki$kokugo)
```

Output

```
[1]  0 35 53 70 96
```

Some descriptives statistics

Input

```
# Mean:
mean(seiseki$kokugo)
```

Output

```
[1] 52.33133
```

Input

```
# Median:
median(seiseki$kokugo)
```

Output

```
[1] 53
```

Input

```
# Minimum, quartiles (Hinge method), maximum:
fivenum(seiseki$kokugo)
```

Output

```
[1]  0 35 53 70 96
```

Tip:
You can reuse R commands that you typed previously by using the arrow keys (↑, ↓) in the Terminal.

Exercise (2)

1. Import the `seiseki.csv` data file into R.
   Save the data as an R object called `seiseki`.

2. Check the variable names in `seiseki` and extract variable `sugaku`.

3. Which value of variable `sugaku` is larger: the mean or the median?

4. Calculate the interquartile range for variable `kokugo` by using the `fivenum()` function.

Exercise (2) — ANSWER

1. Import the `seiseki.csv` data file into R.
   Save the data as an R object called `seiseki`.

Answer

Input

```
seiseki ← read.csv("seiseki.csv")
```

Exercise (2) — ANSWER

2. Check the variable names in `seiseki` and extract variable `sugaku`.

Answer

Input

```
colnames(seiseki)
```

Output

```
 [1] "ID"     "kokugo" "shakai"
 [4] "sugaku" "rika"   "ongaku"
 [7] "bijutu" "taiiku" "gika"
[10] "eigo"
```

Exercise (2) — ANSWER

2. Check the variable names in `seiseki` and extract variable `sugaku`.

Answer

Input

```
colnames(seiseki)
```

Output

```
 [1] "ID"      "kokugo" "shakai"
 [4] "sugaku" "rika"    "ongaku"
 [7] "bijutu" "taiiku" "gika"
[10] "eigo"
```

Input

```
seiseki$sugaku
# or: seiseki[, "sugaku"]
# or: seiseki[, 4]
```

Output

```
 [1] 51  49  23  77  78  56  44  37  7
[10] 29  80  61  36  20  58  22  23  85
[19] 0   24  64  75  86  18  33  59  72
[28] 72  63  38  28  40  50  65  17  62
[37] 91  71  87  59  18  36  85  25  75
[46] 45  35  94  79  84  57  64  29  1
[55] 28  29  48  20  58  94  78  80  91
[64] 69  50  50  39  65  42  47  29  14
[73] 26  23  84  23  39  36  33  36  ...
```

Exercise (2) — ANSWER

3. Which value of variable `sugaku` is larger: the mean or the median?

Answer

Input

```
mean(seiseki$sugaku)
```

Output

```
[1] 45.61446
```

Exercise (2) — ANSWER

3. Which value of variable `sugaku` is larger: the mean or the median?

Answer

Input

```
mean(seiseki$sugaku)
```

Output

```
[1] 45.61446
```

Input

```
median(seiseki$sugaku)
```

Output

```
[1] 43
```

We conclude that the mean is larger than the median.

Exercise (2) — ANSWER

4. Calculate the interquartile range for variable `kokugo` by using the `fivenum()` function.

Answer

Input

```
summ.kokugo ← fivenum(seiseki$kokugo)
summ.kokugo
```

Output

```
[1]  0 35 53 70 96
```

Exercise (2) — ANSWER

4. Calculate the interquartile range for variable `kokugo` by using the `fivenum()` function.

Answer

Input

```
summ.kokugo ← fivenum(seiseki$kokugo)
summ.kokugo
```

Output

```
[1]  0 35 53 70 96
```

Input

```
summ.kokugo[4] - summ.kokugo[2]
```

Output

```
[1] 35
```

The interquartile range of `kokugo` is 35.

Drawing plots

Histograms

Input

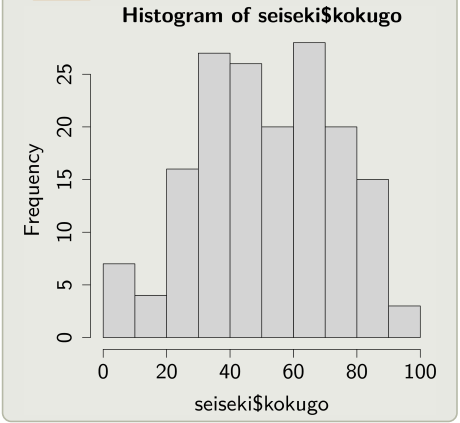
```
# Histogram of the scores of
#  variable 'kokugo':
hist(seiseki$kokugo)
```

Histograms

Input

```
# Histogram of the scores of
#  variable 'kokugo':
hist(seiseki$kokugo)
```

Output



Histograms — Prettifying the plot

The default histogram is a little ugly...
Luckily, we can change any part of the plot to our linking!
In this course, editing plots is optional.

Histograms — Prettifying the plot

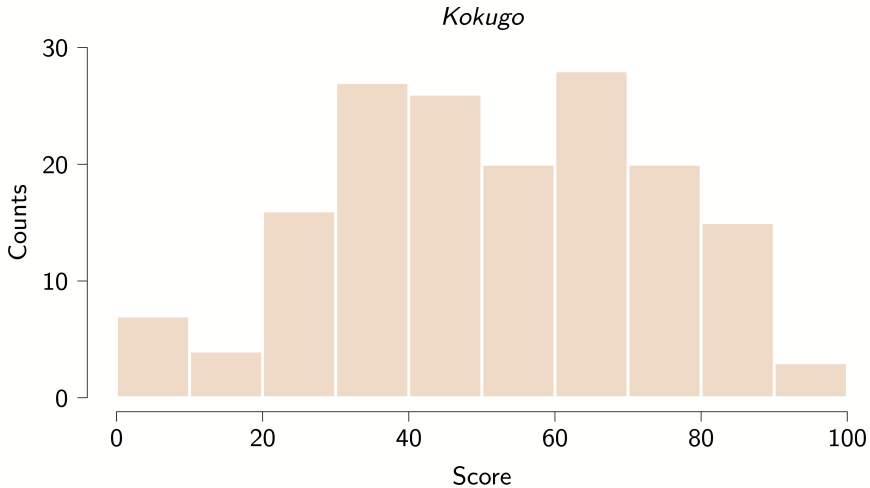
The default histogram is a little ugly...
Luckily, we can change any part of the plot to our linking!
In this course, editing plots is optional.

Input

```
hist(seiseki$kokugo,          # data
     ylim   = c(0, 30),       # y-axis values between 0 and 30
     yaxt   = "n",            # don't draw y-axis (see below)
     col    = "#EDD9C6",      # fill color
     border = "white",        # border color
     xlab   = "Score",        # x-axis label
     ylab   = "Counts",       # y-axis label
     main   = "Kokugo"        # main title
    )

# Custom y-axis:
axis(side = 2,                # 1 = below (x); 2 = left (y); 3 = top; 4 = right
     at   = seq(0, 30, 10),   # position of ticks (0 to 30, in steps of 10)
     las  = 1                 # y axis labels horizontally
    )
```

Histograms (with a few extra tweaks...)



Histograms — Saving to file

You can save the plot as an image file.
That allows you to use the file elsewhere (e.g., in PowerPoint, Word, LaTeX, etc.).

Histograms — Saving to file

You can save the plot as an image file.
That allows you to use the file elsewhere (e.g., in PowerPoint, Word, LaTeX, etc.).

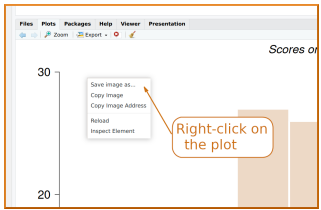
There are multiple ways to do this.

Histograms — Saving to file

You can save the plot as an image file.
That allows you to use the file elsewhere (e.g., in PowerPoint, Word, LaTeX, etc.).

There are multiple ways to do this.

1. Right-click:

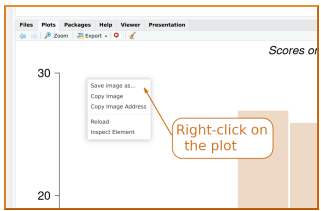


Histograms — Saving to file

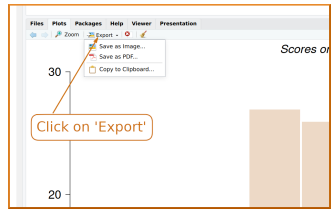
You can save the plot as an image file.
That allows you to use the file elsewhere (e.g., in PowerPoint, Word, LaTeX, etc.).

There are multiple ways to do this.

1. Right-click:



2. Export:



Histograms — Saving to file

You can save the plot as an image file.
That allows you to use the file elsewhere (e.g., in PowerPoint, Word, LaTeX, etc.).

There are multiple ways to do this.

3. R code:

Input

```
png("myfilename.png",
    width     = 4,
    height    = 4,
    units     = "in",
    res       = 300,
    pointsize = 8
   )

hist(seiseki$kokugo)

dev.off()
```

Histograms — Saving to file

You can save the plot as an image file.
That allows you to use the file elsewhere (e.g., in PowerPoint, Word, LaTeX, etc.).

There are multiple ways to do this.

3. R code:

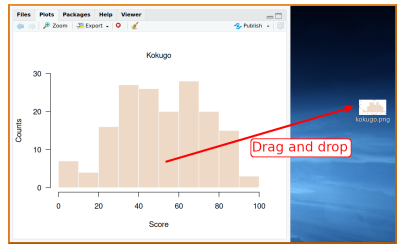
Input

```
png("myfilename.png",
    width     = 4,
    height    = 4,
    units     = "in",
    res       = 300,
    pointsize = 8
   )

hist(seiseki$kokugo)

dev.off()
```

4. Drag and drop:

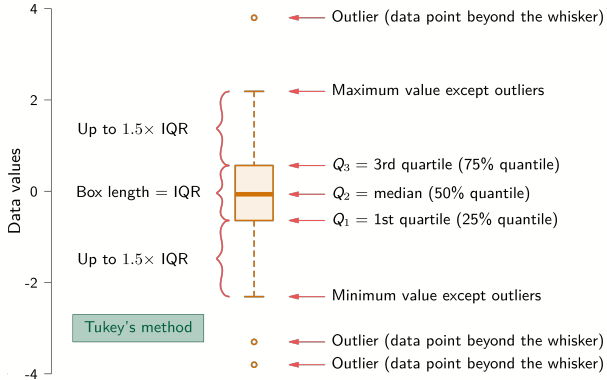


Boxplot (Review from Lecture 04)

Boxplots are plots from which we can easily infer the location and spread of the data!
The bow length gives an indication of the spread of the data.

Boxplot (Review from Lecture 04)

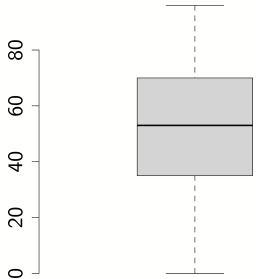
Boxplots are plots from which we can easily infer the location and spread of the data!
The bow length gives an indication of the spread of the data.



Boxplot

Input

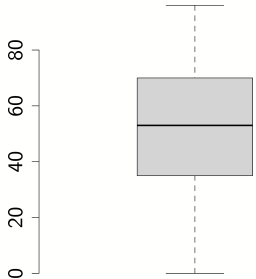
```
# Boxplot for 'kokugo':
boxplot(seiseki$kokugo)
```



Boxplot

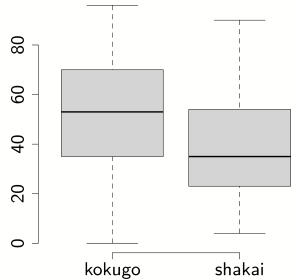
Input

```
# Boxplot for 'kokugo':
boxplot(seiseki$kokugo)
```



Input

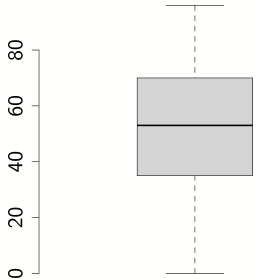
```
# Boxplots side by side
#  (two variables):
boxplot(seiseki$kokugo,
        seiseki$shakai)
```



Boxplot

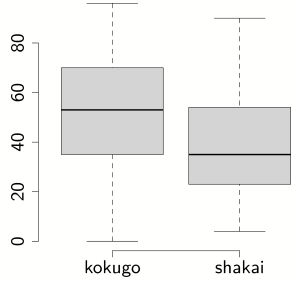
Input

```
# Boxplot for 'kokugo':
boxplot(seiseki$kokugo)
```



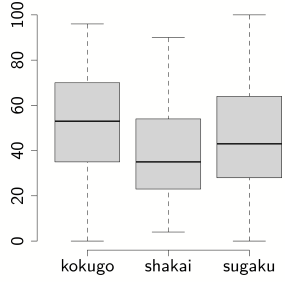
Input

```
# Boxplots side by side
#  (two variables):
boxplot(seiseki$kokugo,
        seiseki$shakai)
```



Input

```
# Boxplots side by side
#  (three variables):
boxplot(seiseki$kokugo,
        seiseki$shakai,
        seiseki$sugaku)
```



Exercise (3)

1. Draw histograms for variables `rika` (Science) and `eigo` (English).
2. From left to right, draw side-by-side boxplots for `sugaku` (Math), `eigo` (English), and `gika` (industrial arts and homemaking), in this order.

From the boxplots, which test subject's data seems the most scattered?

Exercise (3) — ANSWER

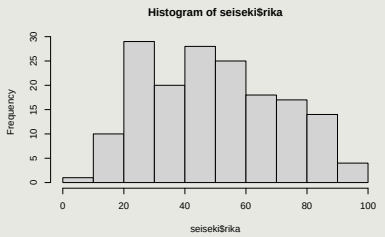
1. Draw histograms for variables `rika` (Science) and `eigo` (English).

Answer

Input

```
hist(seiseki$rika)
```

Output



Exercise (3) — ANSWER

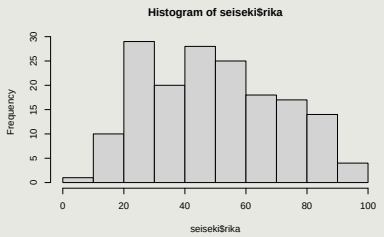
1. Draw histograms for variables `rika` (Science) and `eigo` (English).

Answer

Input

```
hist(seiseki$rika)
```

Output



Input

```
hist(seiseki$eigo)
```

Output

Exercise (3) — ANSWER

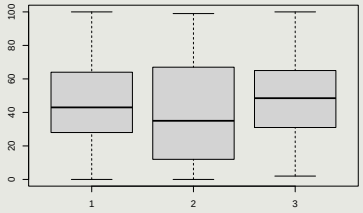
2. From left to right, draw side-by-side boxplots for `sugaku` (Math), `eigo` (English), and `gika` (industrial arts and homemaking), in this order.

Answer

Input

```
boxplot(seiseki$sugaku,
        seiseki$eigo,
        seiseki$gika)
```

Output



Exercise (3) — ANSWER

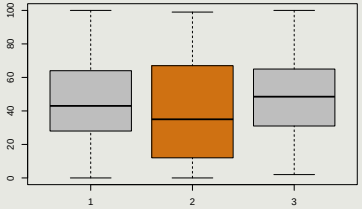
From the boxplots, which test subject's data seems the most scattered?

Answer

Input

```
boxplot(seiseki$sugaku,
        seiseki$eigo,
        seiseki$gika)
```

Output



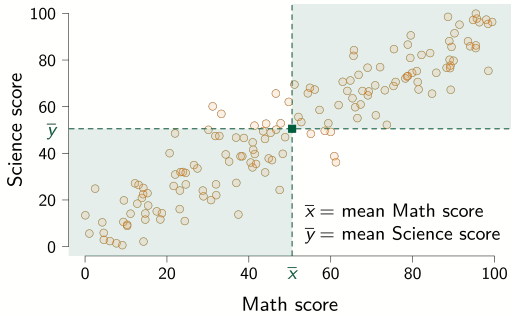
English (eigo) is the most scattered (largest spread).

Scatterplot (Review from Lecture 04)

A scatterplot allows visualizing associations between two variables.

If most points lie in the…

- …top-right and bottom-left panels $\Longrightarrow$ positive linear association.
- …top-left and bottom-right panels $\Longrightarrow$ negative linear association.



Scatterplot

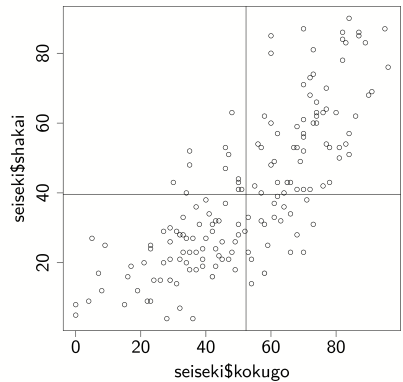
Input

```
# Scatterplot of kokugo (x-axis)
#  vs shakai (y-axis):
plot(seiseki$kokugo, seiseki$shakai)

# AFTER creating the scatterplot, run this:

# Add vertical line at the mean 'kokugo' score:
abline(v = mean(seiseki$kokugo)) # v = vertical

# Add horizontal line at the mean 'shakai' score:
abline(h = mean(seiseki$shakai)) # h = horizontal
```



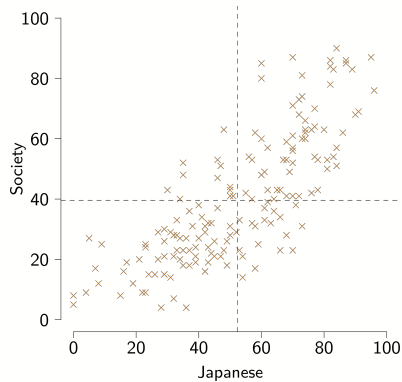
Scatterplot after makeup (completely optional!)

Input

```
# Scatterplot of kokugo (x-axis)
#  vs shakai (y-axis):
plot(seiseki$kokugo, seiseki$shakai,
     pch  = 4,           # points symbol
     col  = "#CF7112",   # points color
     xlab = "Japanese", # x-axis label
     ylab = "Society",  # y-axis label
     xlim = c(0, 100),  # x-axis limits
     ylim = c(0, 100),  # y-axis limits
     xaxt = "n",        # no y axis (below)
     yaxt = "n"         # no y axis (below)
     )

axis(1, seq(0, 100, 20))           # x-axis
axis(2, seq(0, 100, 20), las = 1)  # y-axis

abline(v   = mean(seiseki$kokugo), # vertical
       lty = 2)                    # dashed
abline(h   = mean(seiseki$shakai), # horizontal
       lty = 2)                    # dashed
```



Exercise (4)

1. Draw the following two scatterplots and add lines on the plot, indicating the mean scores.
   1.1 Scatterplot with `sugaku` as x-axis and `eigo` as y-axis.

   1.2 Scatterplot with `sugaku` as x-axis and `taiiku` as y-axis.

2. Describe the data trend in each scatterplot above.

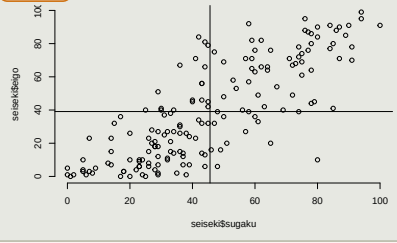
Exercise (4) — ANSWER

Answer

Input

```
# Question 1.1:
plot(seiseki$sugaku, seiseki$eigo)
abline(v = mean(seiseki$sugaku))
abline(h = mean(seiseki$eigo))
```

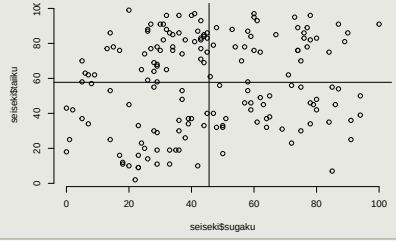
Output



Input

```
# Question 1.2:
plot(seiseki$sugaku, seiseki$taiiku)
abline(v = mean(seiseki$sugaku))
abline(h = mean(seiseki$taiiku))
```

Output

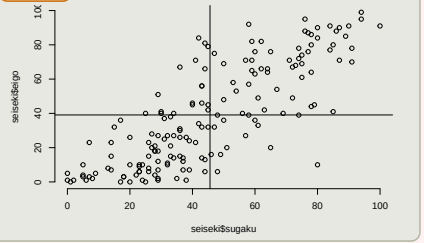


Exercise (4) — ANSWER

2. Describe the data trend in each scatterplot above.

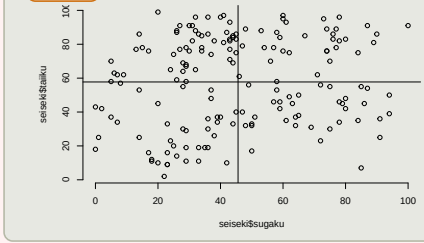
Answer

Output



There is a positive linear association between 'sugaku' and 'eigo'.

Output



There seems to be no linear association between 'sugaku' and 'taiiku'.

Summary

Today we learned how to use R via Posit Cloud (and also locally).

- Import data.
- Compute some descriptive statistics (location, spread).
- Draw some plots (histograms, boxplots, scatterplots).
  Edit features of plots.