

Introduction to R for Research

RStudio
and

R Programming



Workshop Housekeeping



Questions?

Use the chat or raise hand feature.



Feedback Survey

After each session, please fill out the Qualtrics survey (link in chat and emailed out after session).



Credits

Many images are sourced from the teaching team at Harvard Chan Bioinformatics Core (HBC).

Content was similarly inspired by HBC.

Slides with `(HBC Source)` in the bottom corner indicate the image/table source.

Original source: <https://hbctraining.github.io/Training-modules/IntroR/>

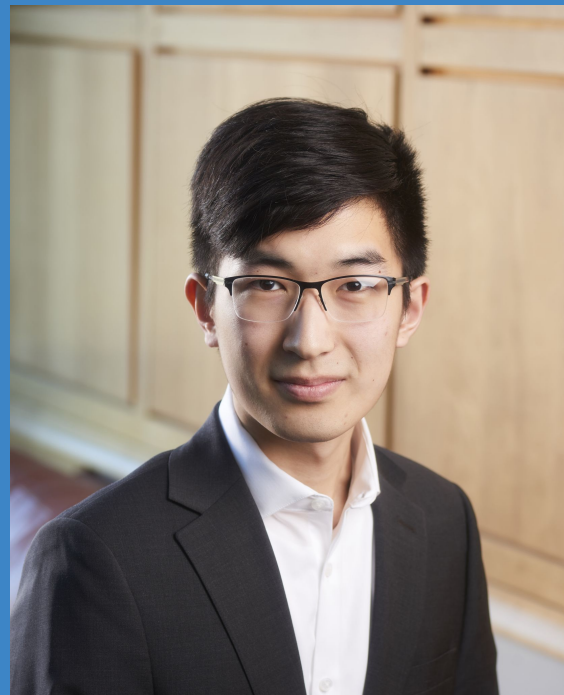
About Me

- Master's in Applied Statistics
- Bachelor's in Computational Statistics

R Experience:

- Self-taught for research in bioinformatics (2017)
- Statistics Courses (STAT 184/380)
- Research data processing (2017-Present)

Research Consultant, University Libraries



David Chen

dzc89@psu.edu

Research Consultant
github.com/TheDavidChen

Introduce Yourself!

Name, Major, Year, Favorite Letter



Motivation

Why do we care?

Motivation - Max Number of Rows

Excel: 1,048,576

R: $2^{48} =$

281,474,976,710,656



Workshop Goals

- General data analysis skills
- Research Skills
- For grad school/career/research applications/opportunities
- Apply to current class/thesis
- Data Visualization
- Manage Big Data
- General Knowledge
- R

Agenda

- Why R/RStudio?
- The RStudio Interface
- Running Code
- Variable Assignment
- Data Types
- Data Structures
- Importing Datasets
- Basic Data Exploration





A Comment on Time

Pace vs Content tradeoff

Why R?

- Free/Open Source
- Platform Agnostic
- Reproducibility
- Popular in Research
- Designed for Data Analysis
- Customizable Data Visualizations
- Big Data!



Logo: The R Foundation

RStudio

- Free/Open Source
- Platform agnostic
- Interface for R

Reasons to use
Base R instead:





General Comments

Expect a learning curve. Expect to struggle.

Errors are the best for learning. If they happen:

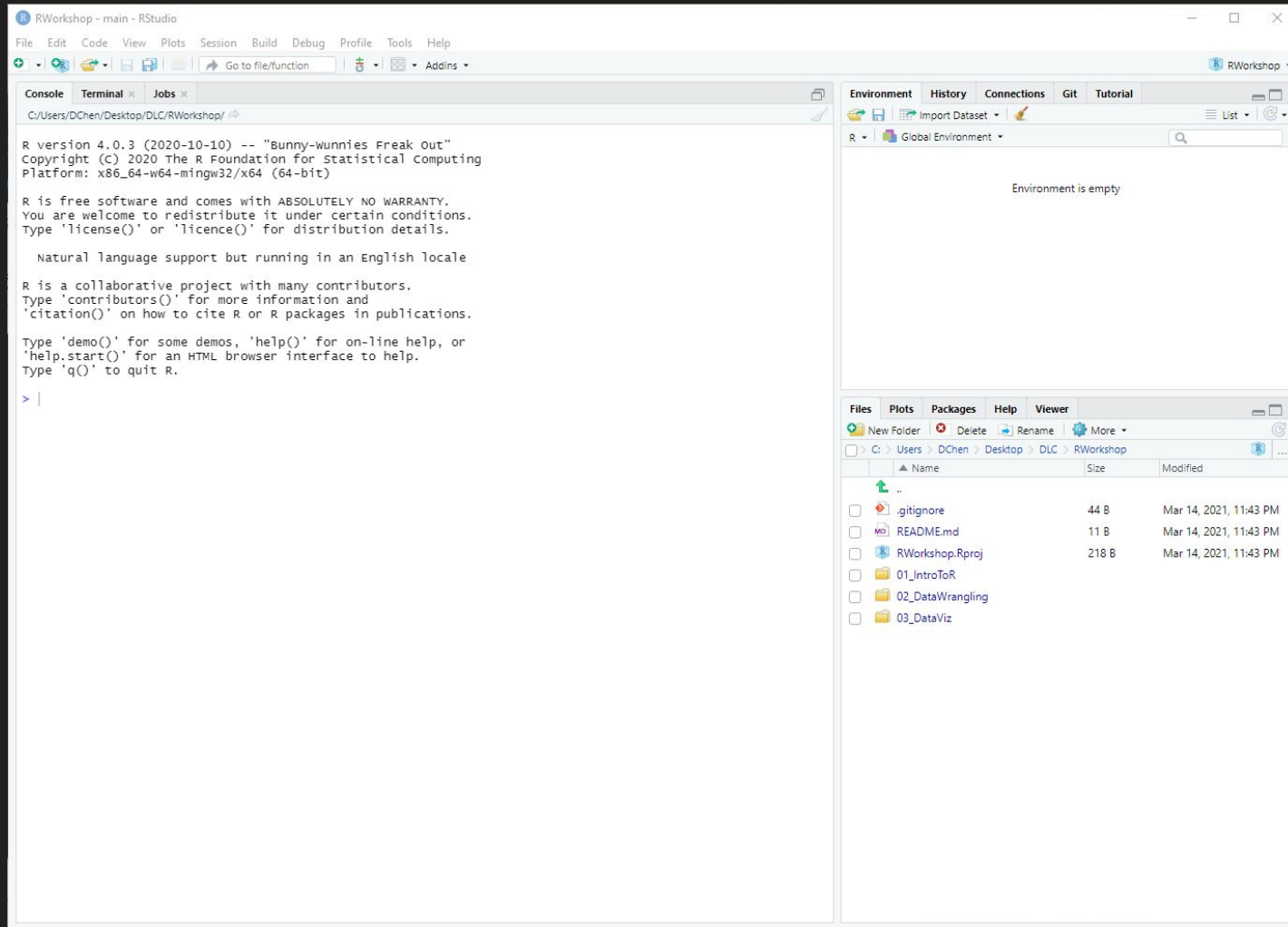
- Read the error and try to solve it
- Share the code that gave you an error
- Share the error message



RStudio is a trademark of RStudio, PBC

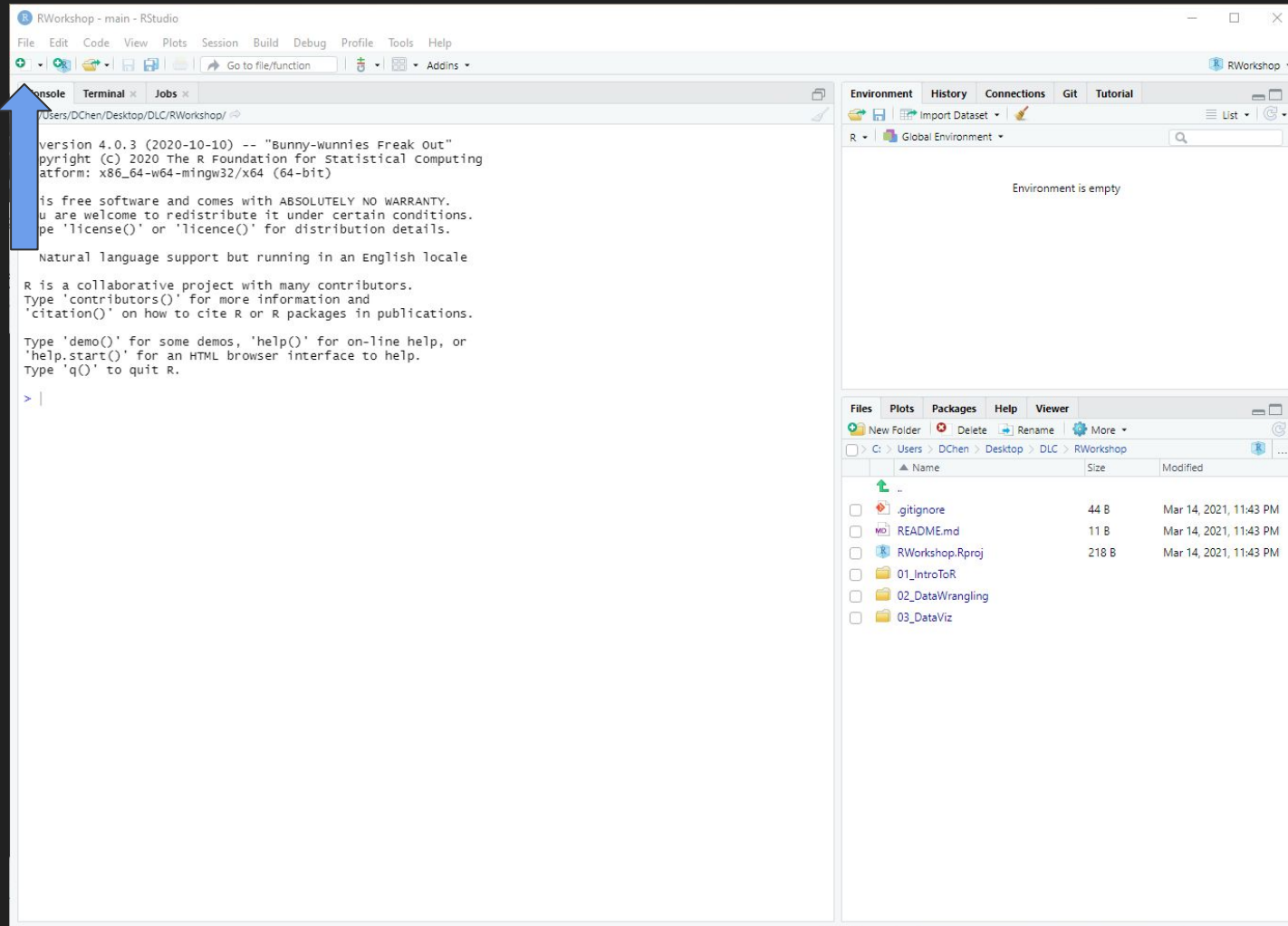
RStudio

Open RStudio!



RStudio

Open RStudio!

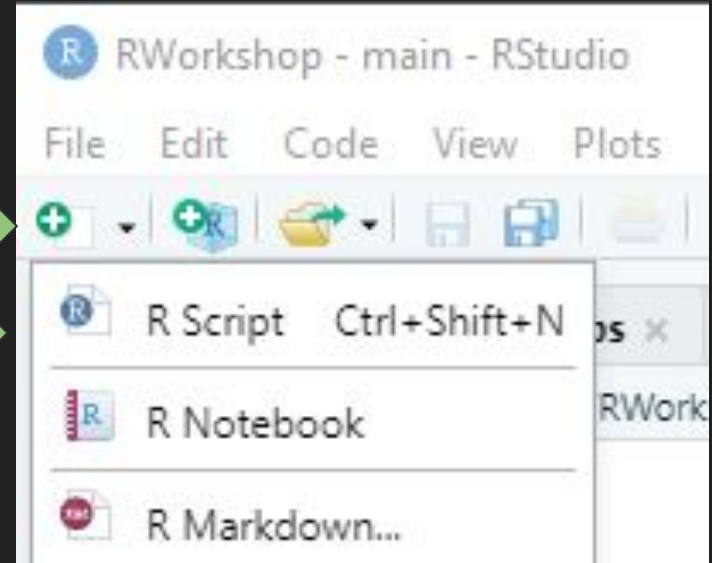


RStudio

Create an R Script

Three options:

- Click the top left button
- Click `File` -> New File -> R Script`
- Press `Ctrl`+`Shift`+`N`



RStudio

The screenshot displays the RStudio IDE interface. The main editor window is titled 'Untitled1' and is currently empty. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The top toolbar contains icons for opening files, saving, running, and other standard IDE functions. The right sidebar is divided into two sections. The top section, labeled 'Environment', shows the 'Global Environment' and indicates that the environment is empty. The bottom section, labeled 'Files', shows the file explorer for the project directory 'C:\Users\DChe\Desktop\DL\RWorkshop'. It lists several files: '..', '.gitignore', 'README.md', 'RWorkshop.Rproj', '01_IntroToR', '02_DataWrangling', and '03_DataViz'. The bottom panel is the 'Console', which shows the R version 4.0.3 (2020-10-10) and the R Foundation for Statistical Computing copyright notice. It also displays the R startup message, including the license information and the natural language support but running in an English locale. The console prompt is '> |'.

R version 4.0.3 (2020-10-10) -- "Bunny-wunnies Freak Out"
Copyright (C) 2020 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.

Natural language support but running in an English locale

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

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

> |

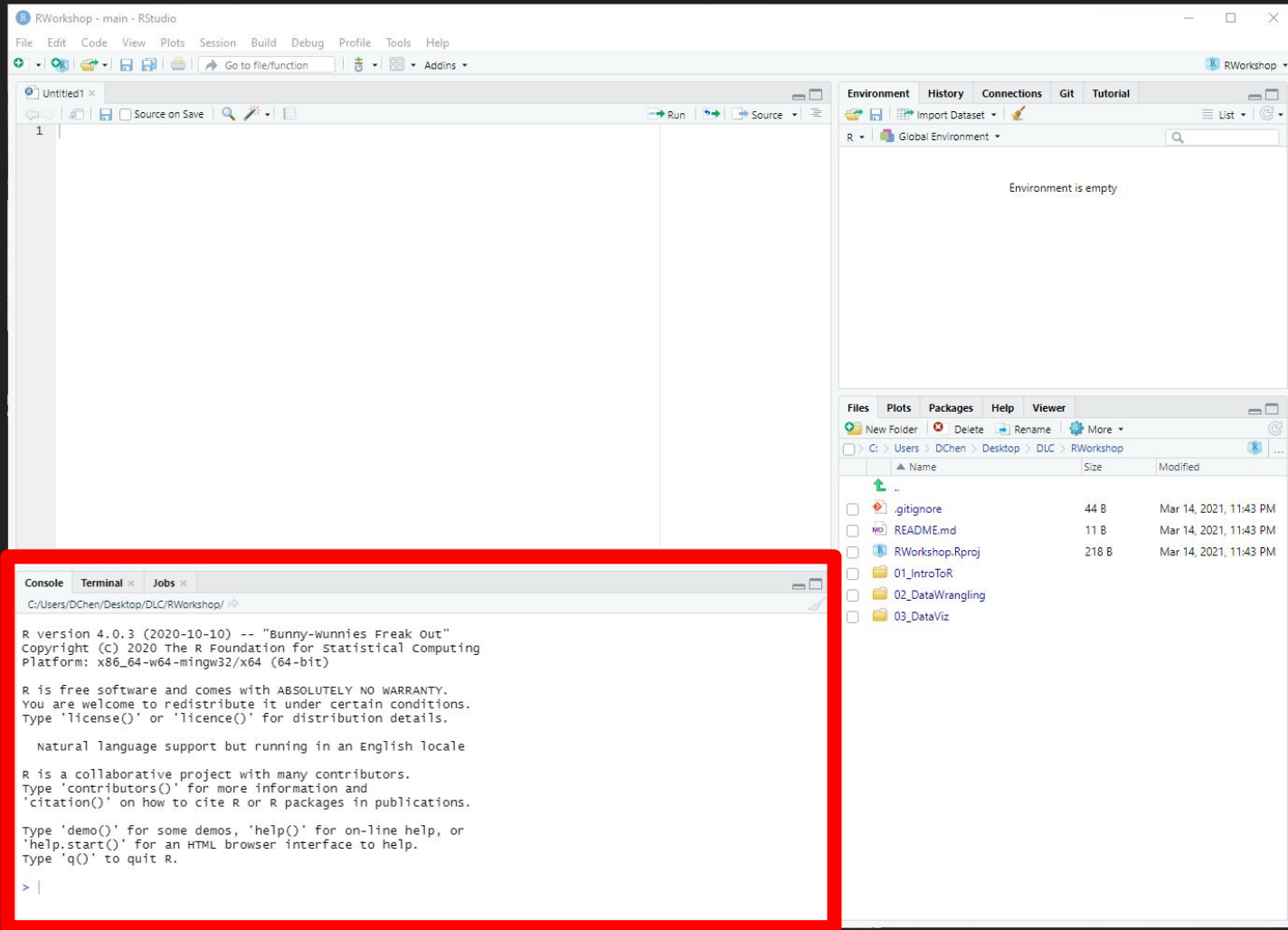
The Console

Run Code

Show Output

Show Errors

Code Unsaved

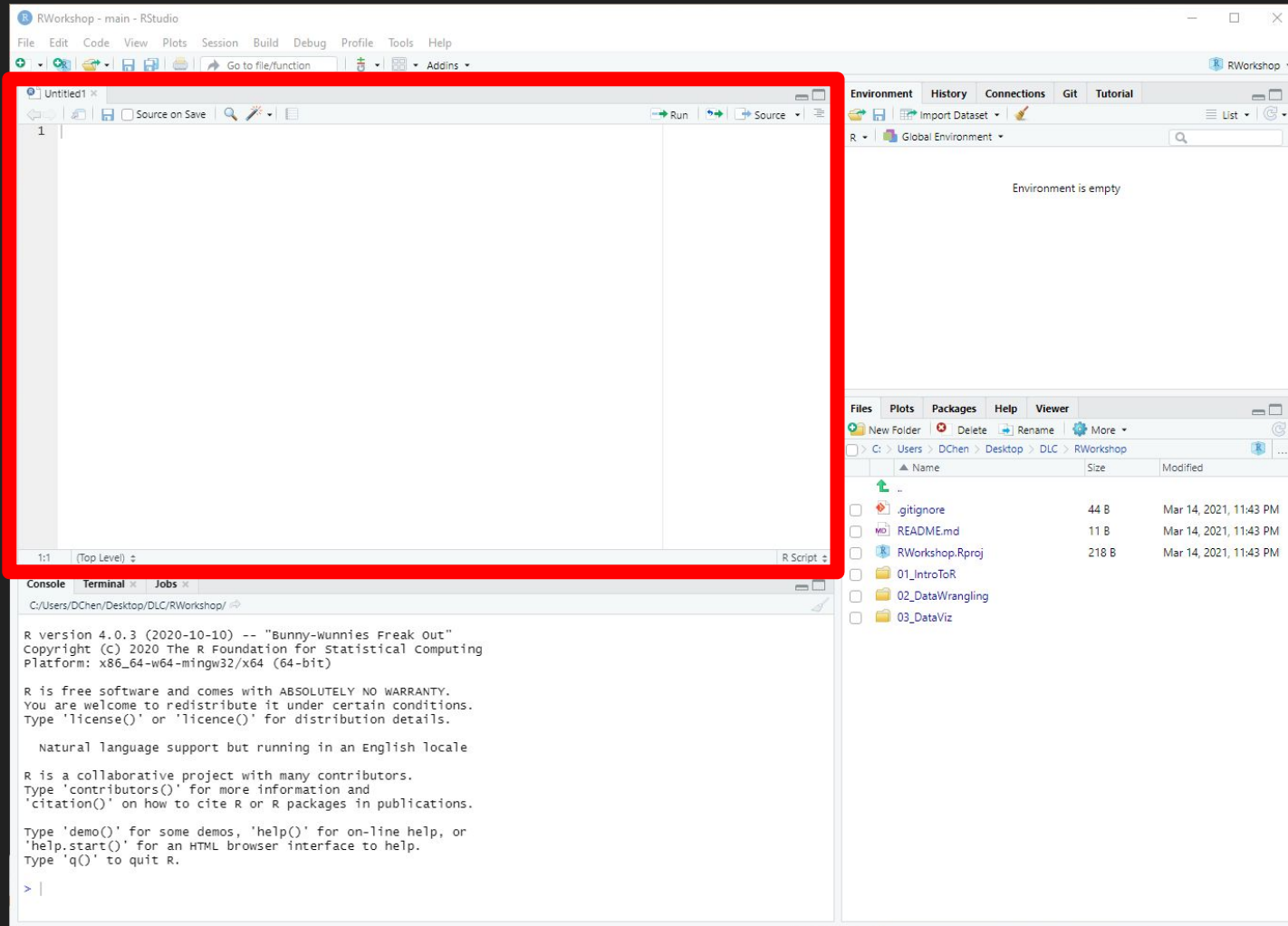


Script Editor

Type code

Run code (output
in console)

Saves all code



Environment

Import Datasets

Shows all
datasets and
defined variables

View imported
datasets

The screenshot displays the RStudio interface. The main window shows an empty R script file named 'Untitled1'. The bottom pane is split into a Console and a Terminal. The Console shows the R version 4.0.3 (2020-10-10) and copyright information. The Terminal shows the R startup message. The right sidebar contains the Environment pane, which is highlighted with a red rectangle. The Environment pane shows the 'Global Environment' and states 'Environment is empty'. Below the Environment pane is a file explorer showing the contents of the 'RWorkshop' directory on the Desktop. The file explorer lists several files and folders, including '.gitignore', 'README.md', 'RWorkshop.Rproj', '01_IntroToR', '02_DataWrangling', and '03_DataViz'.

RStudio - main - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function

Imports Addins

Environment History Connections Git Tutorial

R Global Environment

Environment is empty

1:1 (Top Level) R Script

Console Terminal Jobs

C:/Users/DChen/Desktop/DLC/RWorkshop/

R version 4.0.3 (2020-10-10) -- "Bunny-wunnies Freak Out"
Copyright (C) 2020 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.

Natural language support but running in an English locale

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

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

> |

Files Plots Packages Help View

New Folder Delete Rename More

C:/Users/DChen/Desktop/DLC/RWorkshop

Name	Size	Modified
..		
.gitignore	44 B	Mar 14, 2021, 11:43 PM
README.md	11 B	Mar 14, 2021, 11:43 PM
RWorkshop.Rproj	218 B	Mar 14, 2021, 11:43 PM
01_IntroToR		
02_DataWrangling		
03_DataViz		

Environment

Import Datasets

Shows all
datasets and
defined variables

View imported
datasets

The screenshot shows the RStudio interface with the Environment pane open. The Environment pane displays the Global Environment with a search bar and a list of datasets. The 'Titanic' dataset is selected, showing it has 891 observations and 12 variables. A table of values for the 'x' variable is shown below.

Values
x
5

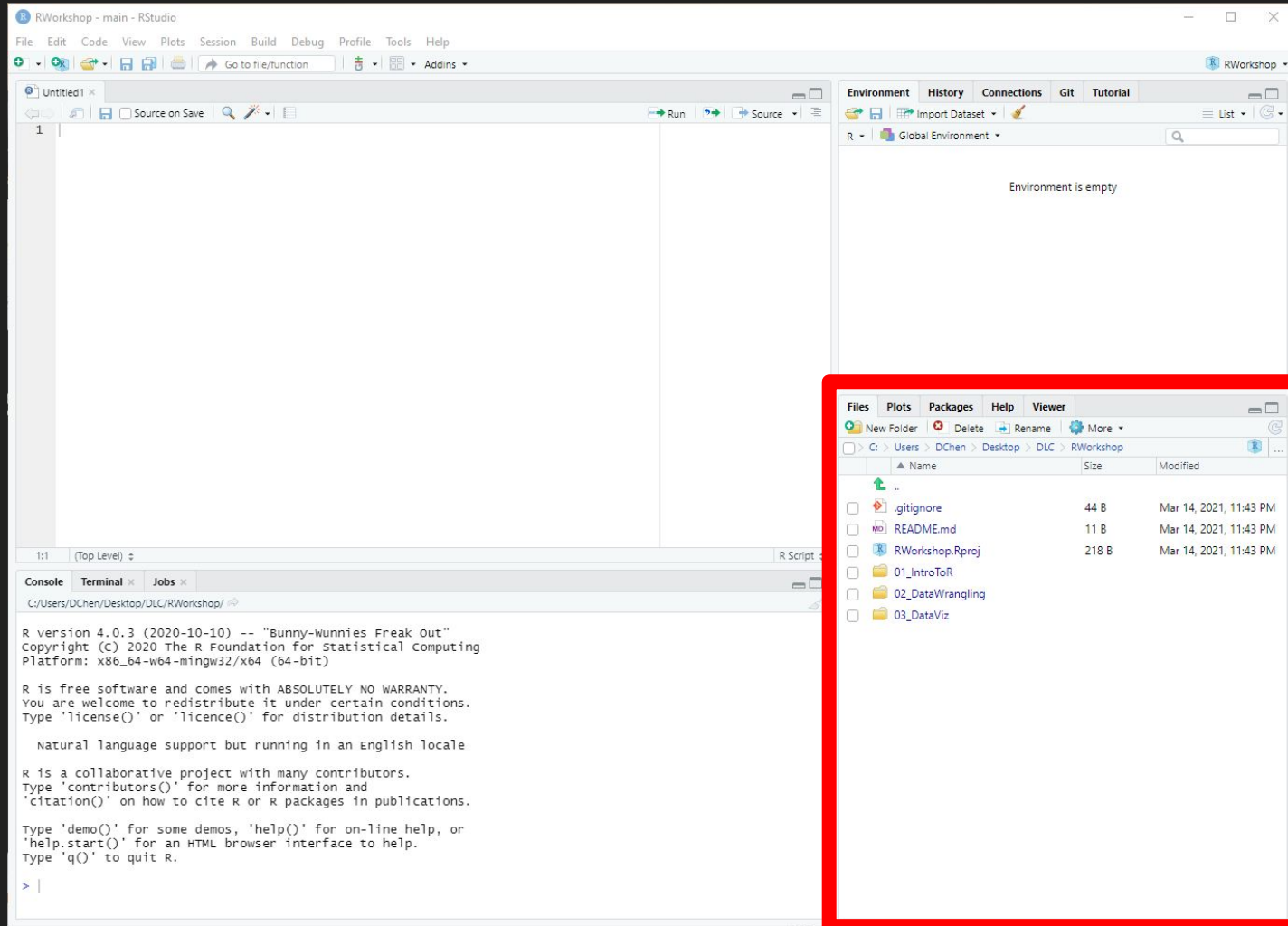
The Files pane on the right shows the project structure, including folders for '01_IntroToR', '02_DataWrangling', and '03_DataViz'.

Misc.

Access Files

View created plots

Read help documentation



The screenshot displays the RStudio environment. The main editor window shows a blank R script file named 'Untitled1'. The console window at the bottom displays the R version 4.0.3 (2020-10-10) -- "Bunny-wunnies Freak Out" copyright notice and instructions on how to use R, including commands like 'license()', 'demo()', 'help()', and 'q()'. A file explorer window is overlaid on the right side of the RStudio interface, showing the directory structure of the RStudio project. The file explorer is titled 'Files' and shows the following files and folders:

Name	Size	Modified
..		
.gitignore	44 B	Mar 14, 2021, 11:43 PM
README.md	11 B	Mar 14, 2021, 11:43 PM
RWorkshop.Rproj	218 B	Mar 14, 2021, 11:43 PM
01_IntroToR		
02_DataWrangling		
03_DataViz		

Running Code

General Comments

If you see: `> print("Hello World")`

Run `print("Hello World")` in R Script. Do not include the ``>``.

Use ``#`` to write comments - code after `#` is not run.

```
> # This is not run
```

Running Code in the Console

Type the following calculations:

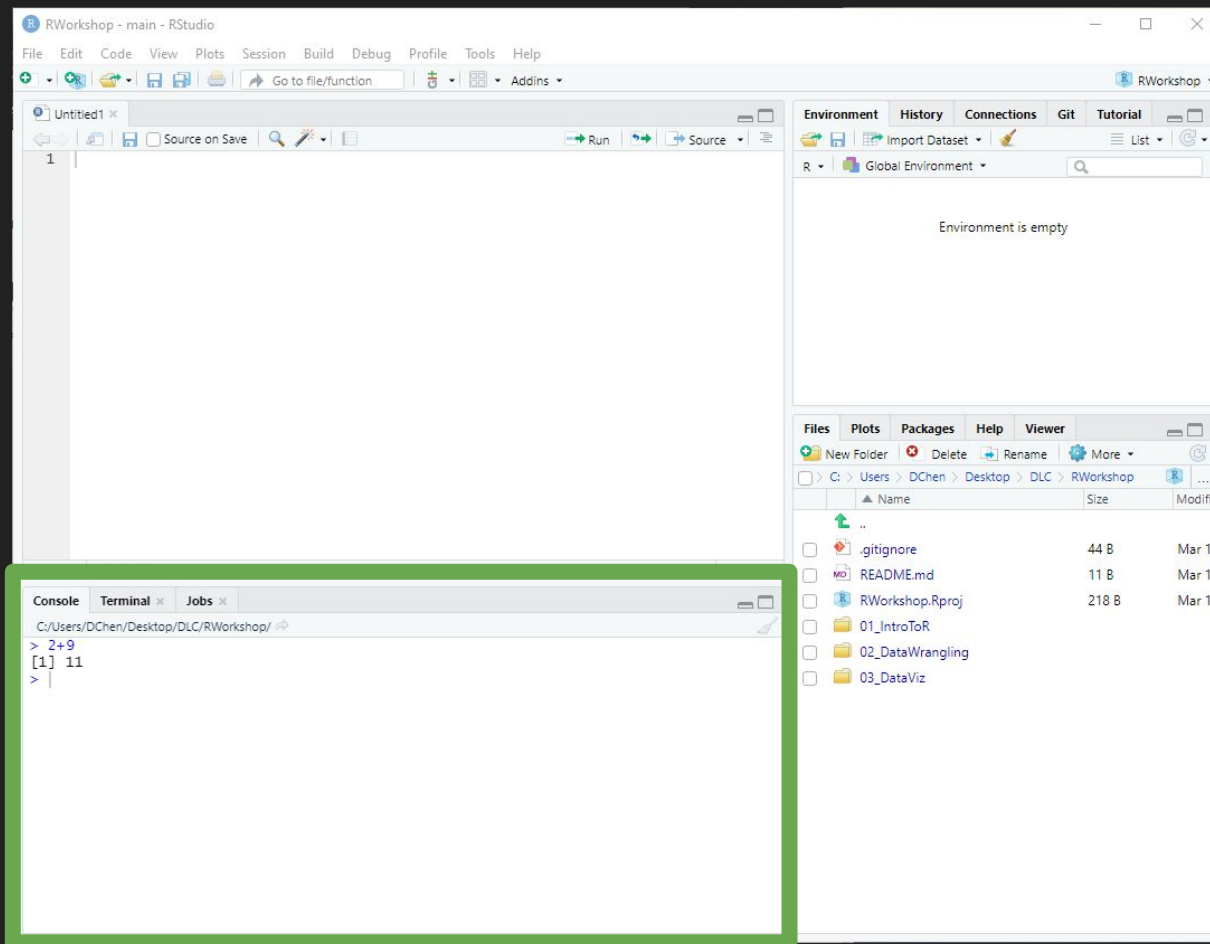
```
> 2+9
```

```
> 15*20
```

```
> 20/5
```

```
> 2^2
```

Press `Enter` to run



Running Code in the Console

Type the following calculations:

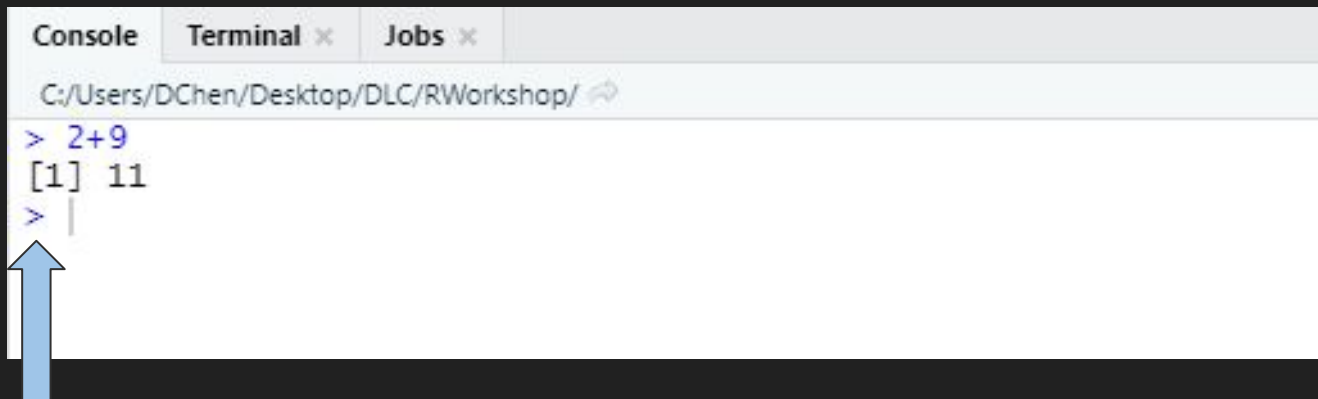
> 2+9

> 15*20

> 20/5

> 2^2

Press `Enter` to run

A screenshot of a console window with tabs for 'Console', 'Terminal', and 'Jobs'. The 'Console' tab is active, showing the path 'C:/Users/DChen/Desktop/DLC/RWorkshop/'. The prompt '>' is followed by the expression '2+9'. The output '[1] 11' is displayed on the next line. A new prompt '>' is on the following line, with a blue arrow pointing to it from below. The arrow originates from the text '>' means it's ready for new code' located below the console window.

```
Console Terminal x Jobs x
C:/Users/DChen/Desktop/DLC/RWorkshop/
> 2+9
[1] 11
> |
```

`>` means it's ready for new code

Running Code in the Console

Type the following calculations:

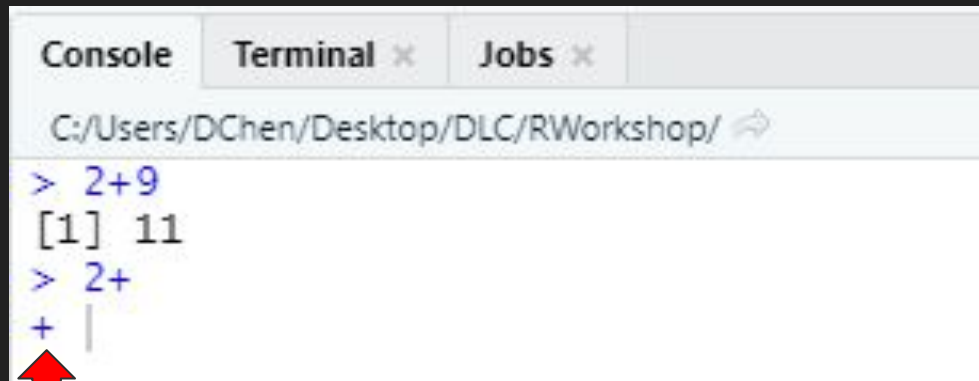
> 2+9

> 15*20

> 20/5

> 2^2

Press `Enter` to run



```
Console Terminal x Jobs x
C:/Users/DChen/Desktop/DLC/RWorkshop/
> 2+9
[1] 11
> 2+
+ |
```

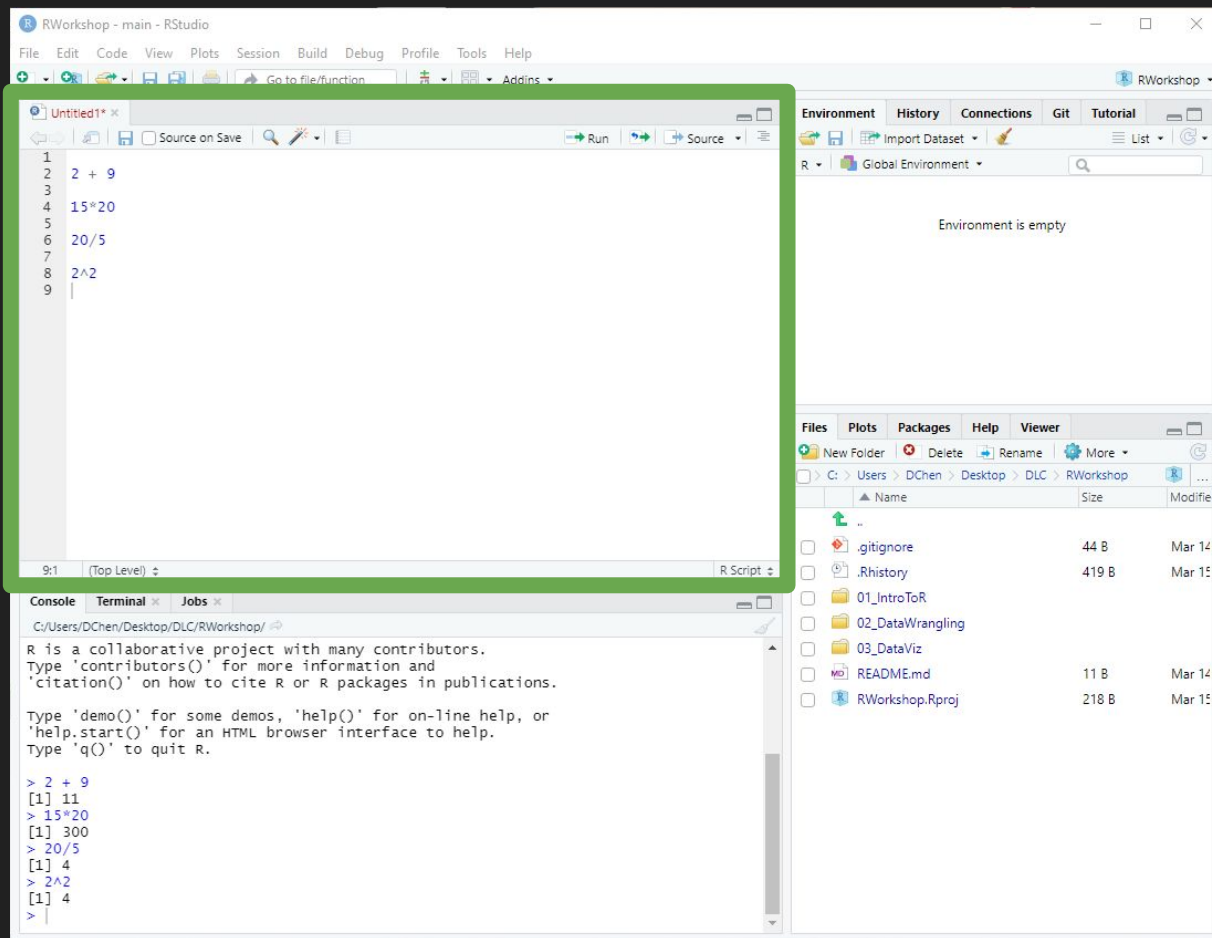
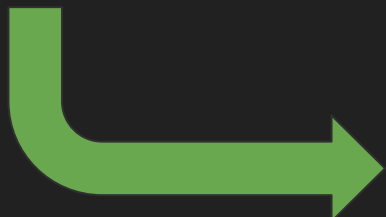
If you see + on the new line:

- Reset with the `Esc` key
- Continue from the previous code

Running Code in the Script

Code in the script can be saved and ran repeatedly

Output shows up in the console

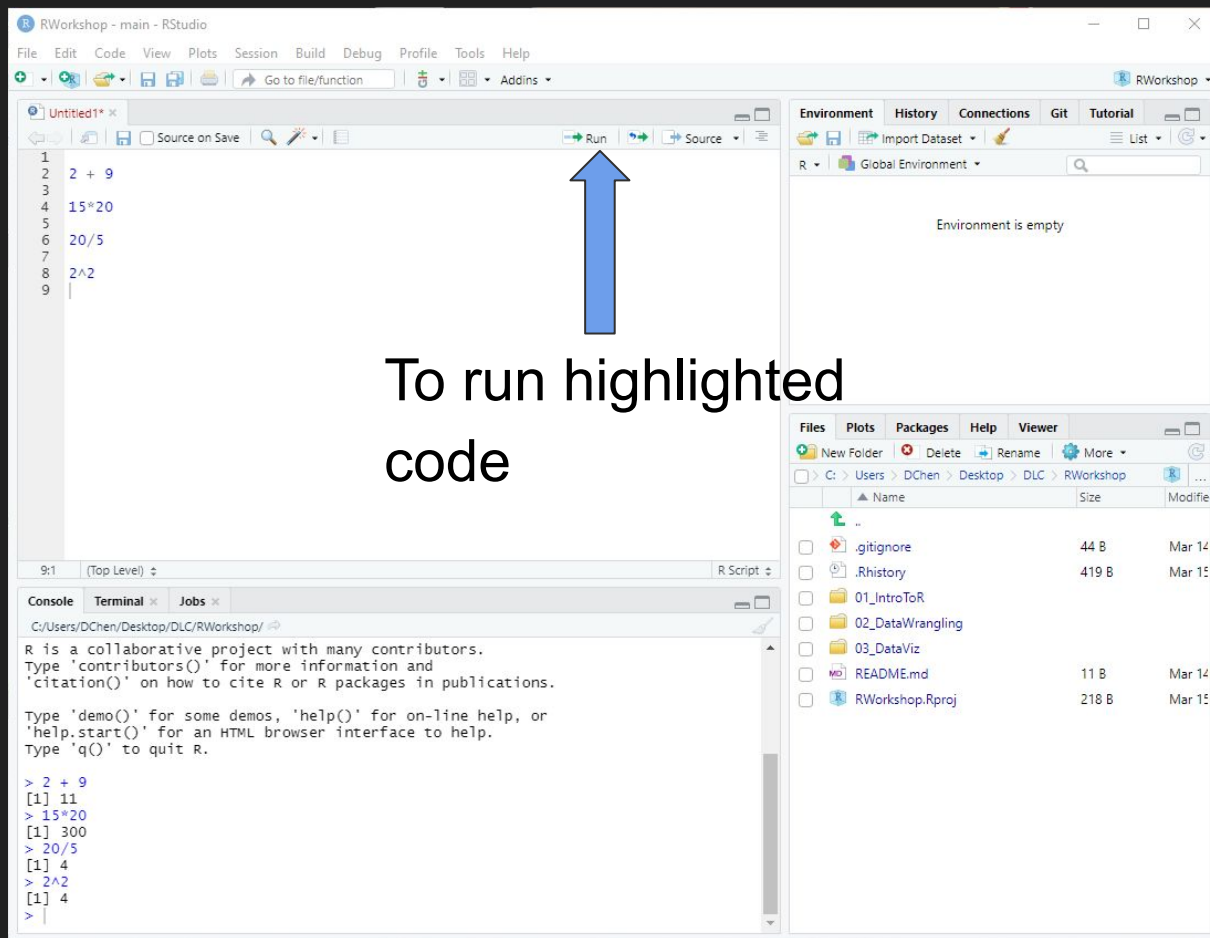


Running Code in the Script

Type code into the R Script

Click or highlight the line of code and

- press `Run`
- press `CTRL` + `ENTER`



To run highlighted code

Running Code in the Script

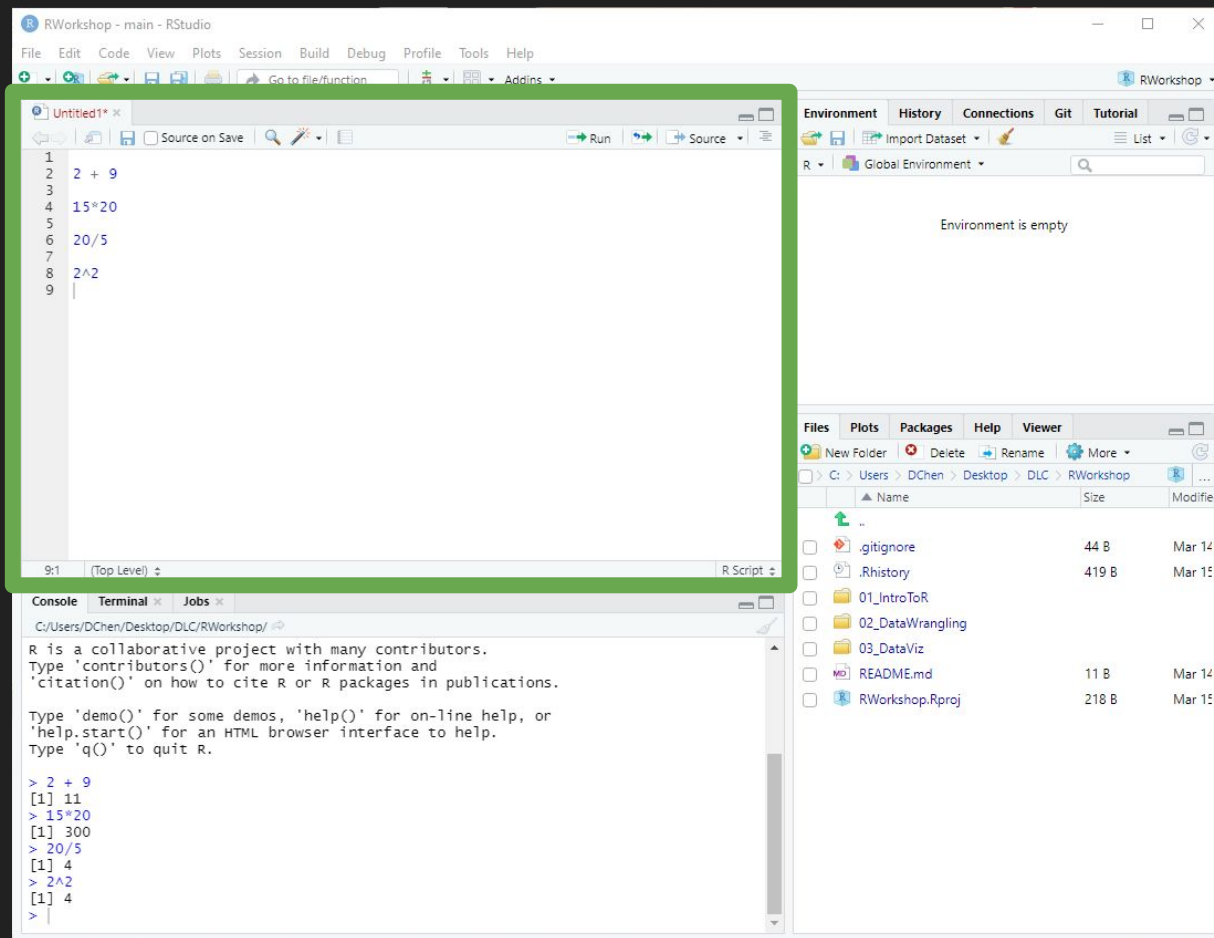
Type the following calculations into the R Script:

```
> 2+9
```

```
> 15*20
```

```
> 20/5
```

```
> 2^2
```



Variable Assignment

variable <- value



Assignment
Operator

(Can also be `=`)

Variable Assignment

variable <- value

Replaces values

```
> x <- 5
```

```
> x + 10
```

Variable Assignment

variable <- value

Replaces values

```
> x <- 5
```

```
> x + 10
```

Redefine over and
over

```
> y <- 5
```

```
> y <- 3
```

```
> y
```

Variable Assignment

variable <- value

Replaces values

```
> x <- 5
```

```
> x + 10
```

Redefine over and
over

```
> y <- 5
```

```
> y <- 3
```

```
> y
```

Multiple
variables

```
> x <- 1
```

```
> y <- 2
```

```
> x * y
```

Variable Assignment - Comments

R is case sensitive

```
> value <- 5
```

```
> VALUE + 5
```

Variable Assignment - Comments

R is case sensitive

```
> value <- 5
```

```
> VALUE + 5
```

Variable names must be
one line and start with a
character

```
> 1x <- 3
```

not valid

```
> x y <- 3
```

not valid

Functions

```
function_name(input)
```

```
> sqrt(4)
```

```
> print("Hello World")
```



Functions

What happens when you try:

```
> sqrt("word")
```

Data Types

Common Data Types

Data Type	Examples
Numeric	-5, 1, 3.33, 100, pi

Common Data Types

Data Type	Examples
Numeric	-5, 1, 3.33, 100, pi
Integer	-5L, 1L, 100L

Common Data Types

Data Type	Examples
Numeric	-5, 1, 3.33, 100, pi
Integer	-5L, 1L, 100L
Character	'words', "3.33", 'TRUE', "1L"

Common Data Types

Data Type	Examples
Numeric	-5, 1, 3.33, 100, pi
Integer	-5L, 1L, 100L
Character	'words', "3.33", 'TRUE', "1L"
Boolean/Logical	TRUE, FALSE, T, F

Common Data Types

Data Type	Examples
Numeric	-5, 1, 3.33, 100, pi
Integer	-5L, 1L, 100L
Character	'words', "3.33", 'TRUE', "1L"
Boolean/Logical	TRUE, FALSE, T, F

Common Data Types - Boolean

Operator	Description
<	Less than
<=	Less than or equal to
>	Greater than
>=	Greater than or equal to
==	Equal to
!=	Not equal to

Common Data Types - Boolean

Examples	Operator	Description
TRUE	<	Less than
<ul style="list-style-type: none">• 5 > 3	<=	Less than or equal to
<ul style="list-style-type: none">• 5 != 3	>	Greater than
FALSE	>=	Greater than or equal to
<ul style="list-style-type: none">• 5 <= 3	==	Equal to
<ul style="list-style-type: none">• 8 == 10	!=	Not equal to

Common Data Types - Boolean

Try it yourself!

```
> 5 > 3
```

```
> 10 == 10
```

Applies to variables!

```
> x <- 5
```

```
> x == 5
```

```
> x > 10
```

Operator

Description

<

Less than

<=

Less than or equal to

>

Greater than

>=

Greater than or
equal to

==

Equal to

!=

Not equal to

Common Data Types

Check the type of a variable/object with ``class()``:

```
> class("4")
```

```
> class(4)
```

Certain functions require specific data types:

```
> sqrt("4")
```

Common Data Types - Common Mistakes

If a number is in quotations, think of it as the word instead of the number.

- “5” != 5

Common Data Types - Common Mistakes

If a number is in quotations, think of it as the word instead of the number.

- `"5" != 5`

TRUE/FALSE are actually encoded as 1/0

- `TRUE == 1 ; FALSE == 0`

Common Data Types - Common Mistakes

If a number is in quotations, think of it as the word instead of the number.

- `"5" != 5`

TRUE/FALSE are actually encoded as 1/0

- `TRUE == 1 ; FALSE == 0`

Characters without quotation marks are variables!

- `Hi <- 5`

Common Data Types - Common Mistakes

If a number is in quotations, think of it as the word instead of the number.

- `"5" != 5`

TRUE/FALSE are actually encoded as 1/0

- `TRUE == 1 ; FALSE == 0`

Characters without quotation marks are variables!

- `Hi <- 5`

Be careful of single `=`; outside of functions, it is assignment!

- `5 = 5 --> error!`
- `x = 5` is equivalent to `x <- 5`, but not recommended

Data Structures

Vectors

1	50	9	42
---	----	---	----

"A"	"B"	"C"
-----	-----	-----

TRUE	F
------	---

A column in excel

Any length from 1 onwards

All values have to be same type (all numeric, boolean, character, etc.)

One value in a vector is called an **element**

Vectors

Create vectors with function `c()`

```
> lengths <- c(50, 100, 150, 200)
```

```
> color <- c("blue", "white", "purple")
```

Bonus: What happens if you try to add a different type?

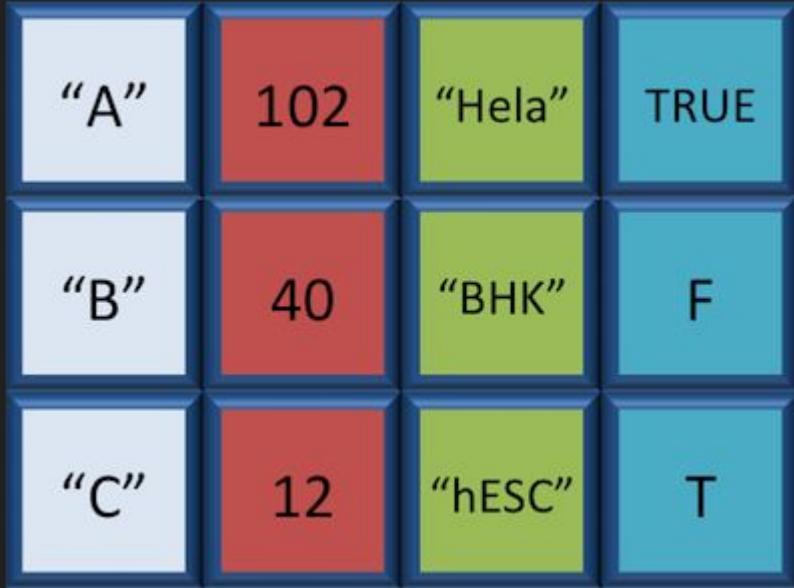
```
> bonus <- c("Yes", TRUE, 5)
```


Factors



A special vector with values assigned to each factor level (category)

Data Frame



A 3x4 grid representing a data frame. The cells are colored as follows: the first column is light blue, the second is red, the third is light green, and the fourth is light blue. Each cell contains a value, and the grid is enclosed in a dark blue border.

"A"	102	"Hela"	TRUE
"B"	40	"BHK"	F
"C"	12	"hESC"	T

Most common data format;
equivalent to an excel sheet.

Multiple vectors combined
together

Columns must be same type

All columns must have equal
number of rows

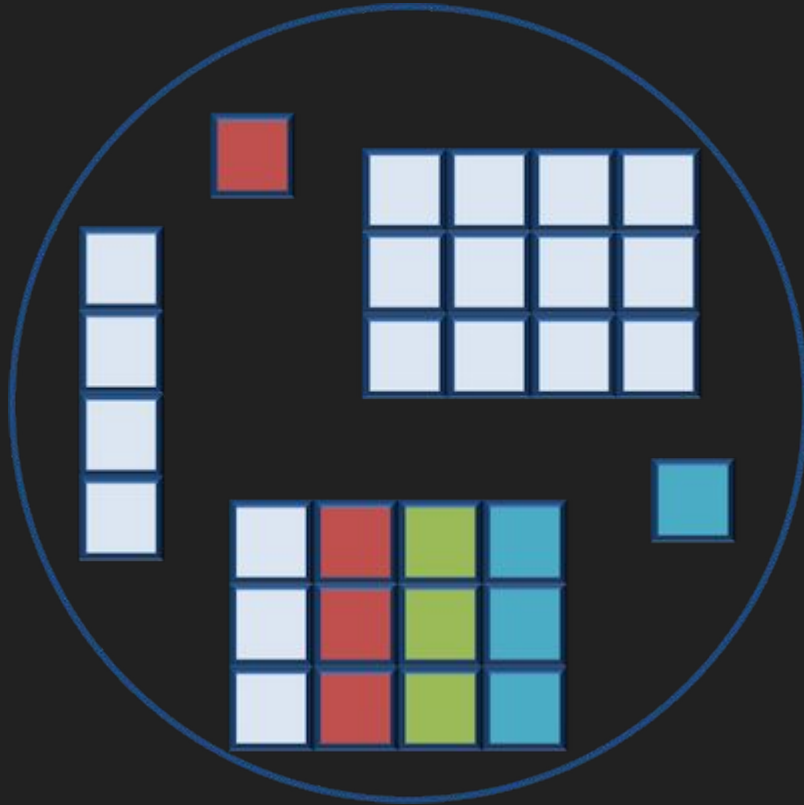
Matrix

90	5	137	9
87	40	2	52
4	102	32	41

Data frame that requires same type and length

Used in some statistical/mathematical functions

Lists



Holds any number of data structures or types

Think of it as the excel file, where it can contain multiple sheets

Reading in Data

Titanic Dataset: <https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/stuff/titanic.csv>

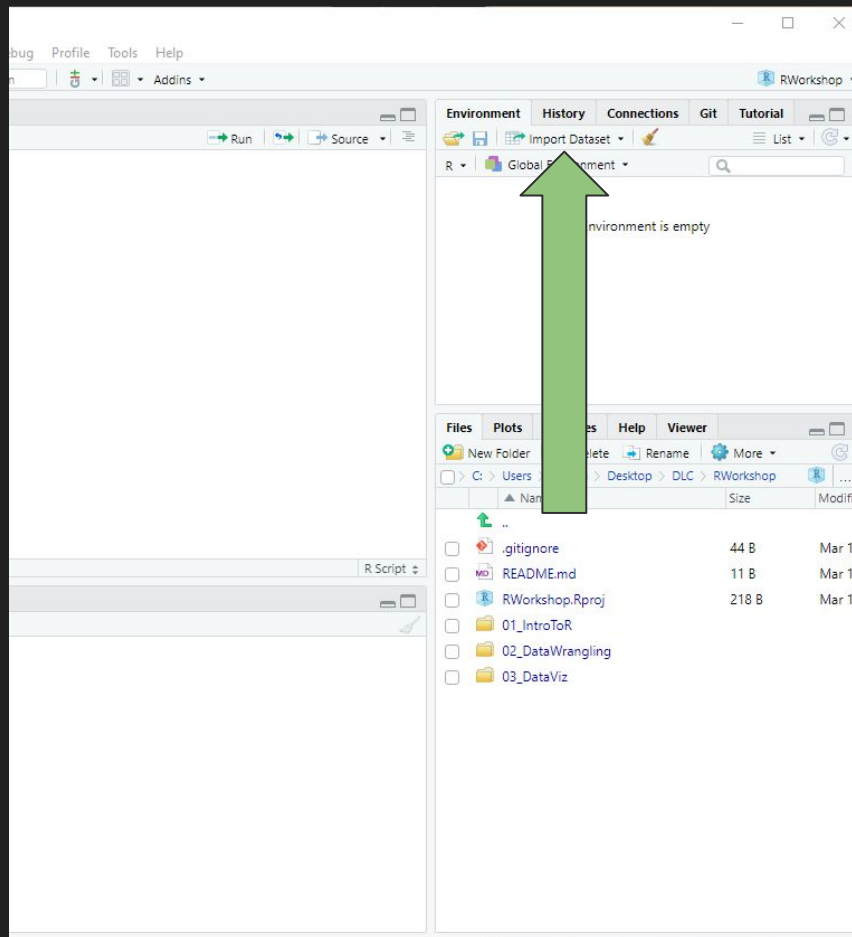
Reading in Datasets

Data type	Extension	Function	Package
Comma separated values	csv	<code>read.csv()</code>	utils (default)
		<code>read_csv()</code>	readr (tidyverse)
Tab separated values	tsv	<code>read_tsv()</code>	readr
Other delimited formats	txt	<code>read.table()</code>	utils
Stata version 7-12	dta	<code>read.dta()</code>	foreign
SPSS	sav	<code>read.spss()</code>	foreign
SAS	sas7bdat	<code>read.sas7bdat()</code>	sas7bdat
Excel	xlsx, xls	<code>read_excel()</code>	readxl (tidyverse)

(HBC Source)

Reading in Datasets

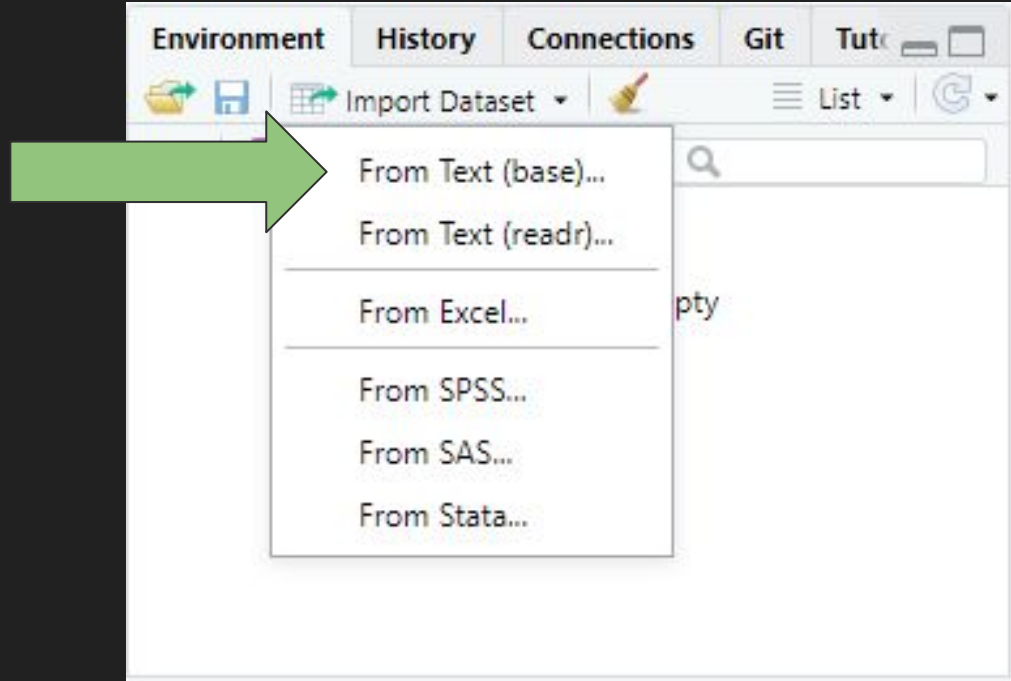
`Import Dataset` in the environment will automatically generate the code for you!



Reading in Datasets

From Text works for almost all basic data files (.csv, .txt, etc.)

Find and select the dataset



Reading in Datasets

Import Dataset

Name
titanic

Encoding
Automatic

Heading
☒ Yes ☐ No

Row names
Automatic

Separator
Comma

Decimal
Period

Quote
Double quote (")

Comment
None

na.strings
NA

☐ Strings as factors

Input File

Survived,Pclass,Name,Sex,Age,Siblings/Spouses Aboard,Parent
0,3,Mr. Owen Harris Braund,male,22,1,0,7.25
1,1,Mrs. John Bradley (Florence Briggs Thayer) Cumings,fem
1,3,Miss. Laina Heikkinen,female,26,0,0,7.925
1,1,Mrs. Jacques Heath (Lily May Peel) Futrelle,female,35,:
0,3,Mr. William Henry Allen,male,35,0,0,8.05
0,3,Mr. James Moran,male,27,0,0,8.4583
0,1,Mr. Timothy J McCarthy,male,54,0,0,51.8625
0,3,Master. Gosta Leonard Palsson,male,2,3,1,21.075
1,3,Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson,femal
1,2,Mrs. Nicholas (Adele Achem) Nasser,female,14,1,0,30.07
1,3,Miss. Marguerite Rut Sandstrom,female,4,1,1,16.7
1,1,Miss. Elizabeth Bonnell,female,58,0,0,26.55
0,3,Mr. William Henry Saunderson,male,20,0,0,8.05
0,3,Mr. Anders Johan Andersson,male,39,1,5,31.275
0,3,Miss. Hulda Amanda Adolfina Vestrom,female,14,0,0,7.85
1,2,Mrs. (Mary D Kingcome) Hewlett,female,55,0,0,16

Data Frame

Survived	Pclass	Name
0	3	Mr. Owen Harris Braund
1	1	Mrs. John Bradley (Florence Briggs Thayer) Cumings
1	3	Miss. Laina Heikkinen
1	1	Mrs. Jacques Heath (Lily May Peel) Futrelle
0	3	Mr. William Henry Allen
0	3	Mr. James Moran
0	1	Mr. Timothy J McCarthy
0	3	Master. Gosta Leonard Palsson
1	3	Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson
1	2	Mrs. Nicholas (Adele Achem) Nasser
1	3	Miss. Marguerite Rut Sandstrom
1	1	Miss. Elizabeth Bonnell
0	3	Mr. William Henry Saunderson
0	3	Mr. Anders Johan Andersson
0	3	Miss. Hulda Amanda Adolfina Vestrom
1	2	Mrs. (Mary D Kingcome) Hewlett

Import Cancel

Reading in Datasets

Make sure the data is being imported correctly!

Check:

- Headers
- Columns
- Values

Import Dataset

Name: titanic

Encoding: Automatic

Heading: ☒ Yes ☐ No

Row names: Automatic

Separator: Comma

Decimal: Period

Quote: Double quote (")

Comment: None

na.strings: NA

☐ Strings as factors

Input File: Survived,Pclass,Name,Sex,Age,Siblings/Spouses Aboard,Parent...

Data Frame:

Survived	Pclass	Name
0	3	Mr. Owen Harris Braund
1	1	Mrs. John Bradley (Florence Briggs Thayer) Cumings,femi
1	3	Miss. Laina Heikkinen
1	1	Mrs. Jacques Heath (Lily May Peel) Futr
0	3	Mr. William Henry Allen
0	3	Mr. James Moran
0	1	Mr. Timothy J McCarthy
0	3	Master. Gosta Leonard Palsson
1	3	Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson,femal
1	2	Mrs. Nicholas (Adele Achem) Nasser
1	3	Miss. Marguerite Rut Sandstrom
1	1	Miss. Elizabeth Bonnell
0	3	Mr. William Henry Saunderson
0	3	Mr. Anders Johan Andersson
0	3	Miss. Hulda Amanda Adolfin Vestrom
1	2	Mrs. (Mary D Kingcome) Hewlett

Import Cancel

Viewing the Datasets

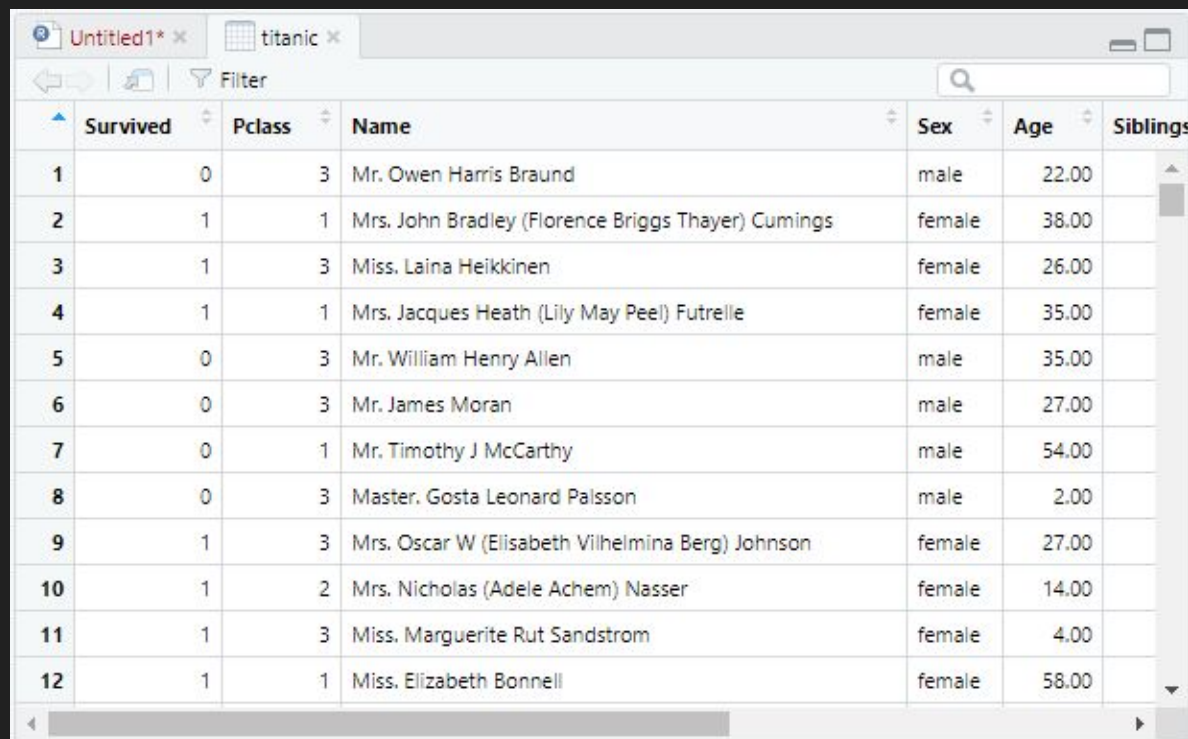
The data is now in the environment!

Click the name (titanic) to view the data!



Viewing the Datasets

The data view after clicking the name in the environment



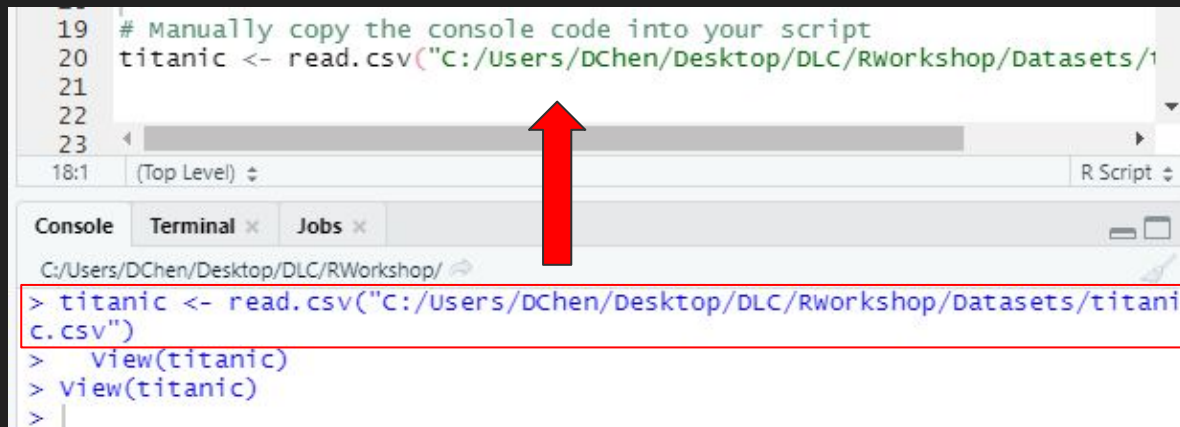
	Survived	Pclass	Name	Sex	Age	Siblings
1	0	3	Mr. Owen Harris Braund	male	22.00	
2	1	1	Mrs. John Bradley (Florence Briggs Thayer) Cumings	female	38.00	
3	1	3	Miss. Laina Heikkinen	female	26.00	
4	1	1	Mrs. Jacques Heath (Lily May Peel) Futrelle	female	35.00	
5	0	3	Mr. William Henry Allen	male	35.00	
6	0	3	Mr. James Moran	male	27.00	
7	0	1	Mr. Timothy J McCarthy	male	54.00	
8	0	3	Master. Gosta Leonard Palsson	male	2.00	
9	1	3	Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson	female	27.00	
10	1	2	Mrs. Nicholas (Adele Achem) Nasser	female	14.00	
11	1	3	Miss. Marguerite Rut Sandstrom	female	4.00	
12	1	1	Miss. Elizabeth Bonnell	female	58.00	

After Importing the Data

Copy the generated code into the R Script.

In the future, you will only need to run the code.

Do not include the `>` or the `View()` functions.



```
19 # Manually copy the console code into your script
20 titanic <- read.csv("C:/Users/DChen/Desktop/DLC/Rworkshop/Datasets/titanic.csv")
21
22
23
```

18:1 (Top Level) R Script

Console Terminal x Jobs x

C:/Users/DChen/Desktop/DLC/Rworkshop/

```
> titanic <- read.csv("C:/Users/DChen/Desktop/DLC/Rworkshop/Datasets/titanic.csv")
> view(titanic)
> view(titanic)
>
```

Examine the Data - Try These Functions!

Functions - Add `titanic` to ()

str()

dim()

summary()

nrow()

head()

ncol()

tail()

colnames()

```
> str(titanic)
```

```
> head(titanic)
```

```
> colnames(titanic)
```

Next Workshop

Data Processing

```
> install.packages("tidyverse")
```