Before we start...

Access the RStudio Cloud Project (Link in chat)

- Workshop - Introduction to R & Data Wrangling

1. Download R
2. Download RStudio
3. Download the `titanic` dataset

   https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/problem12.html

Concept of data analysis and maintenance Illustration by WOOBRO DESIGN on Iconscout

Introduction to R and Data Wrangling

(with tidyverse)

Penn State, University Libraries, Research Informatics and Publishing

Workshop Housekeeping

Questions?

Use the Q&A, chat, or raise hand feature.

Feedback Survey

After the session, please fill out the Qualtrics survey (link in chat).

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 (2017-Present)
- Statistics Courses (STAT 184/380)
- Other Projects

Research Consultant, University Libraries

David Chen
dzc89@psu.edu
Research Consultant
github.com/TheDavidChen

Introduce Yourself!

Name, Major, Year, Favorite Letter

Agenda

Introduction to R

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

Agenda

Introduction to R

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

Data Wrangling with Tidyverse

- Why Tidyverse?
- The Fundamental Functions
  - Subset data
  - Create new variables
  - Grouped summaries
  - Combining datasets
- Exporting Datasets

Why R?

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

Logo: The R Foundation

Motivation - Max Number of Rows

Excel: 1,048,576

R: 2^48 = 281,474,976,710,656

https://support.microsoft.com/en-us/office/excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3
https://cran.r-project.org/doc/manuals/R-ints.html#Long-vectors

RStudio

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

Reasons to use Base R instead:

RStudio is a trademark of RStudio, PBC

A Comment on Time

Pace vs Content tradeoff

General Comments

Expect a learning curve. Expect to struggle.

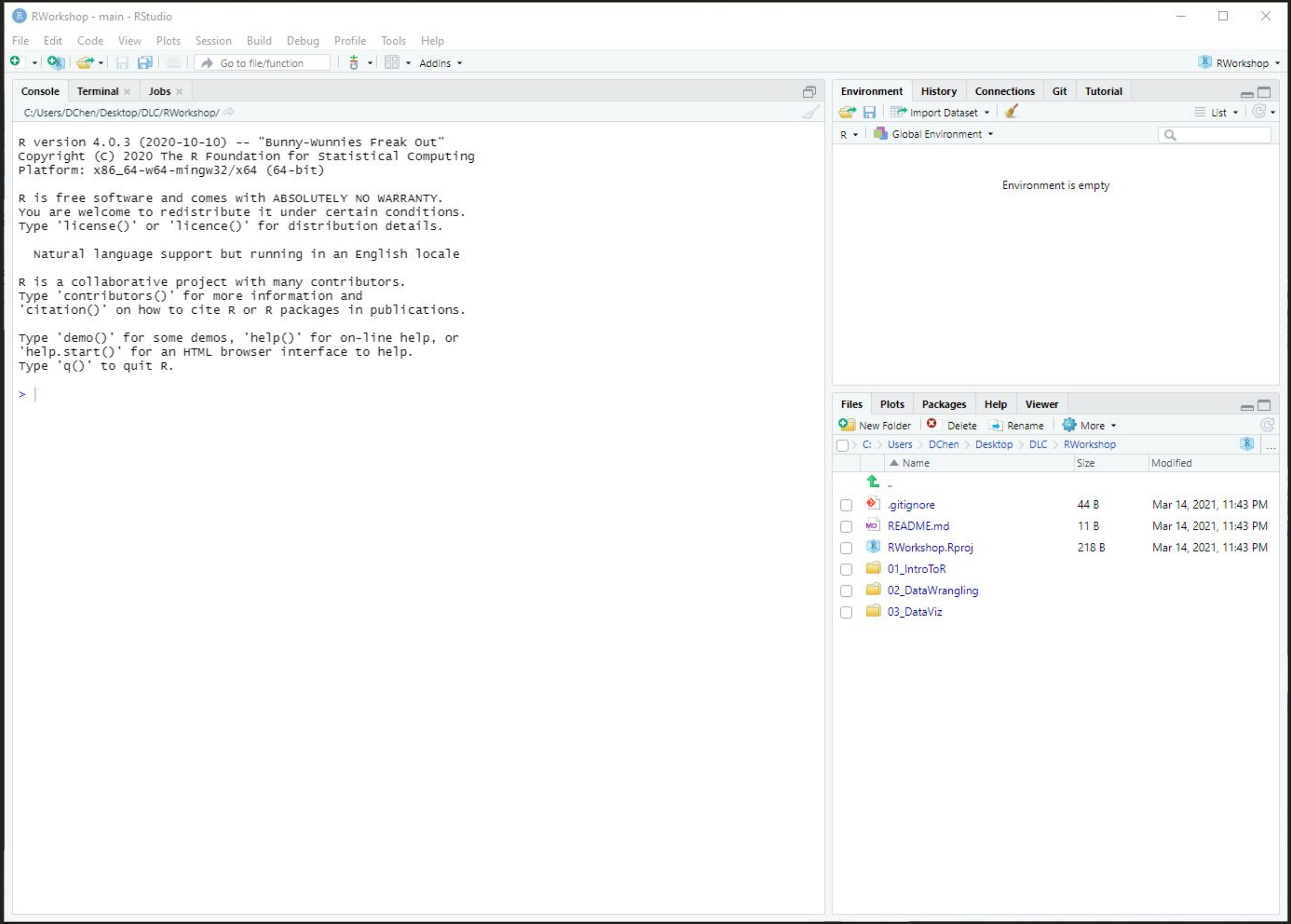
Errors are the best for learning. If they happen:

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

RStudio®

RStudio

Open RStudio!

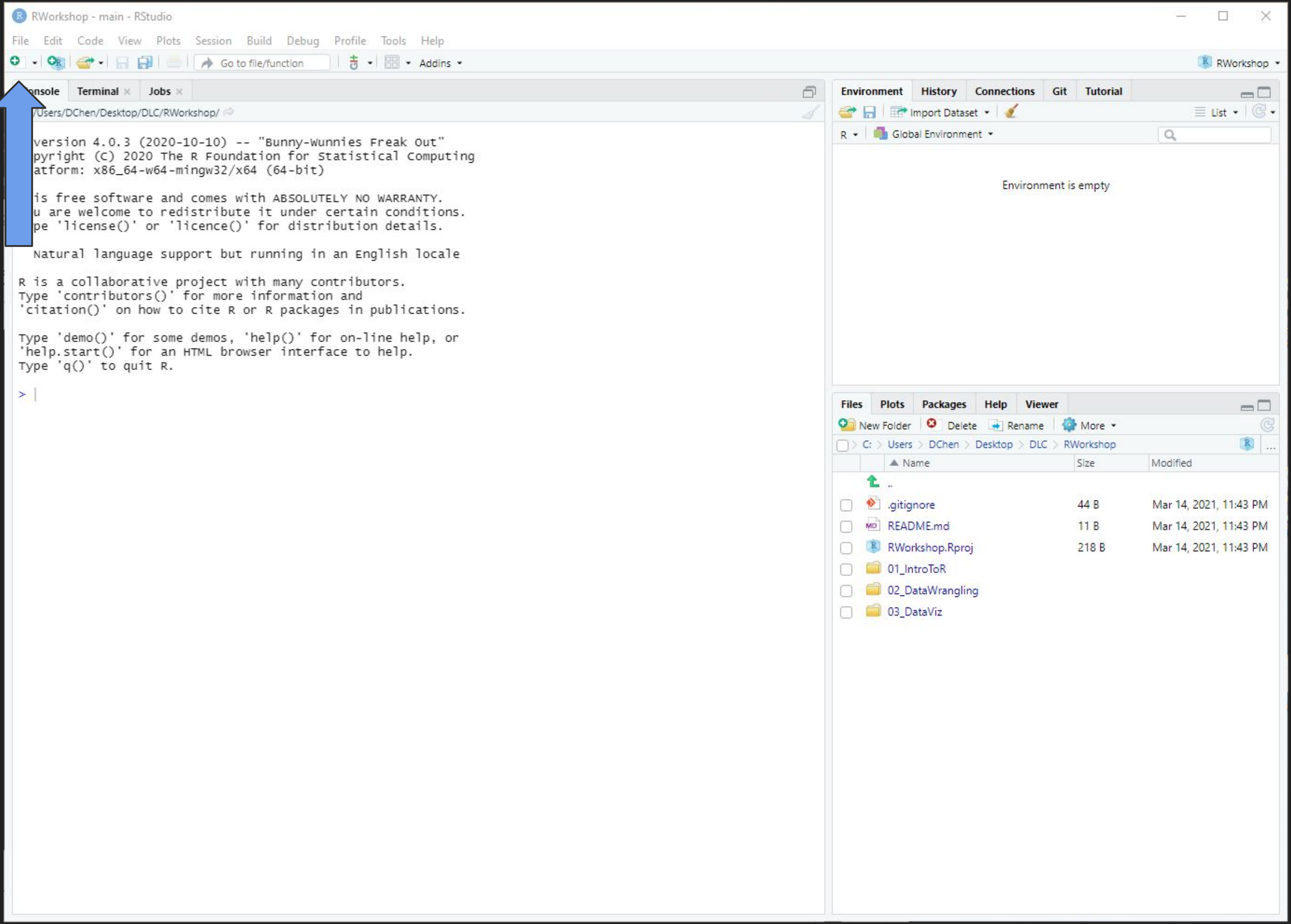

RWorkshop - main - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
RWorkshop
Console Terminal Jobs
C:/Users/DChen/Desktop/DLC/RWorkshop/
Environment History Connections Git Tutorial
Import Dataset
List
R
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: Users DChen Desktop DLC RWorkshop
Name Size Modified

```
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.

> 
```

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		

RStudio

Open RStudio!


RWorkshop - main - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
RWorkshop
nsole Terminal Jobs
Environment History Connections Git Tutorial
Import Dataset
List
R Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: Users DChen Desktop DLC RWorkshop
Name Size Modified

```
/Users/DChen/Desktop/DLC/RWorkshop/

version 4.0.3 (2020-10-10) -- "Bunny-Wunnies Freak Out"
pyright (C) 2020 The R Foundation for Statistical Computing
atform: x86_64-w64-mingw32/x64 (64-bit)

is free software and comes with ABSOLUTELY NO WARRANTY.
u are welcome to redistribute it under certain conditions.
pe '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.

> |
```

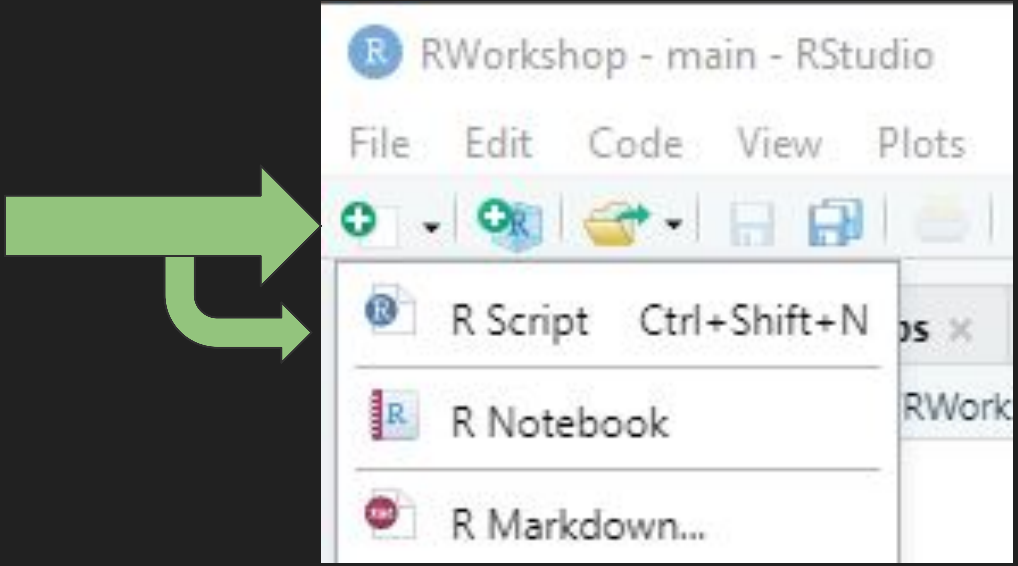
	Name	Size	Modified
	..		
	.gitignore	44 B	Mar 14, 2021, 11:43 PM
	README.md	11 B	Mar 14, 2021, 11:43 PM
	RWorkshop.Rproj	218 B	Mar 14, 2021, 11:43 PM
	01_IntroToR		
	02_DataWrangling		
	03_DataViz		

RStudio

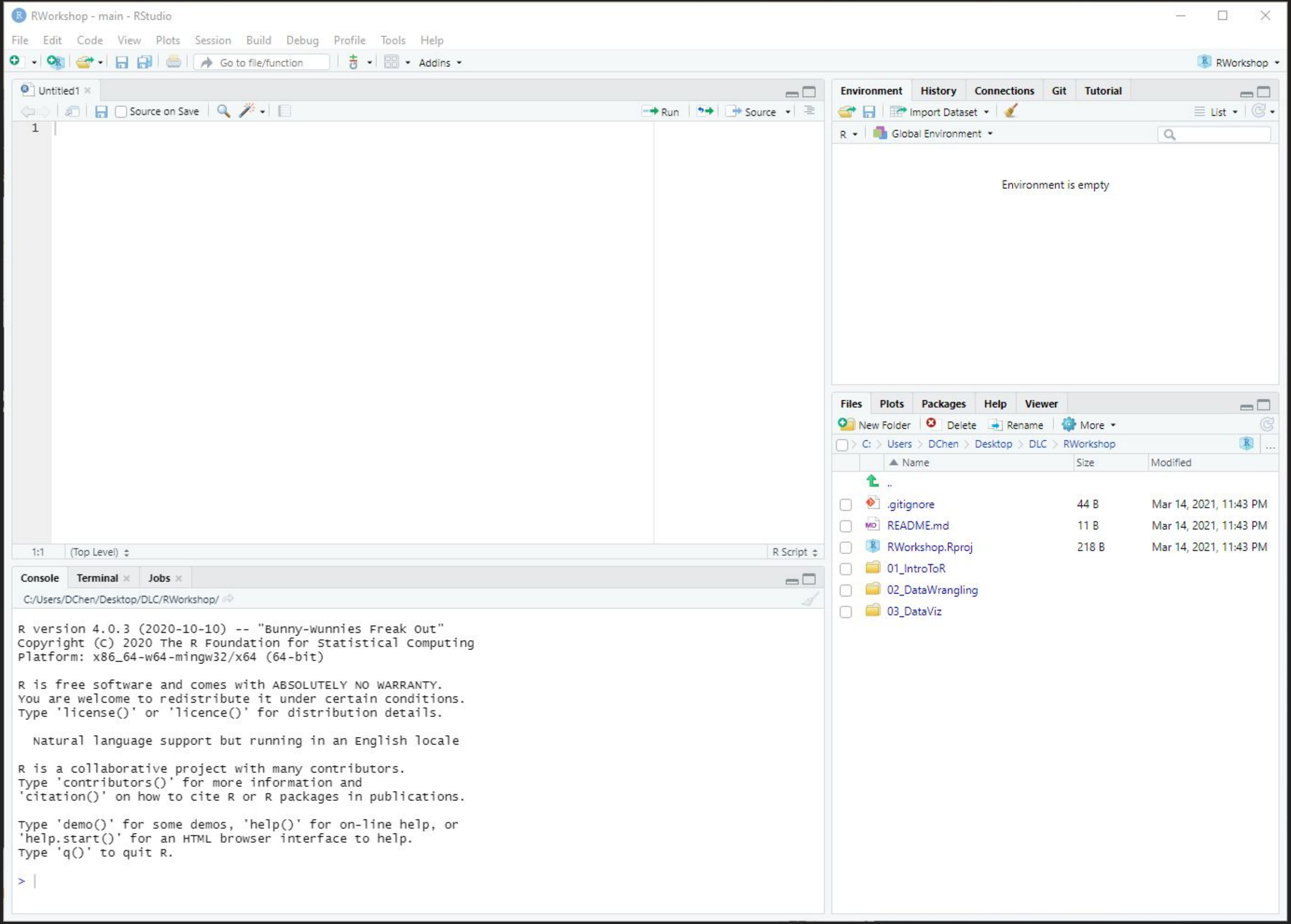
Create an R Script

Three options:

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



RStudio


RWorkshop - main - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
RWorkshop
Untitled1
Source on Save
Run
Source
1:1
(Top Level)
R Script
Environment History Connections Git Tutorial
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: > Users > DChen > Desktop > DLC > RWorkshop
Console Terminal Jobs
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		

```
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.

> 
```

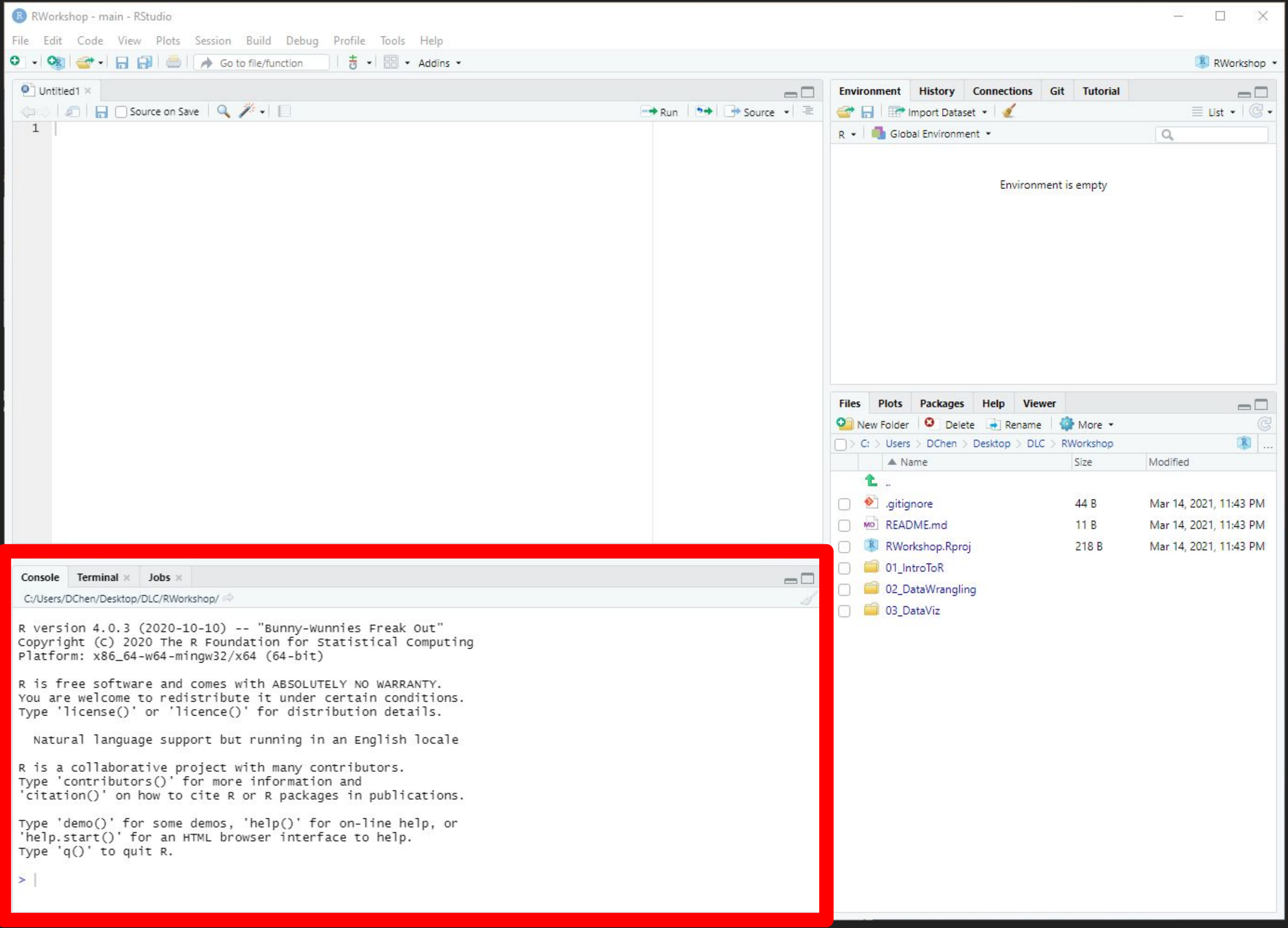
The Console

Run Code

Show Output

Show Errors

Code <u>Unsaved</u>


RWorkshop - main - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
RWorkshop
Untitled1
Source on Save
Run
Source
Environment History Connections Git Tutorial
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
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
Console Terminal Jobs
C:/Users/DChen/Desktop/DLC/RWorkshop/

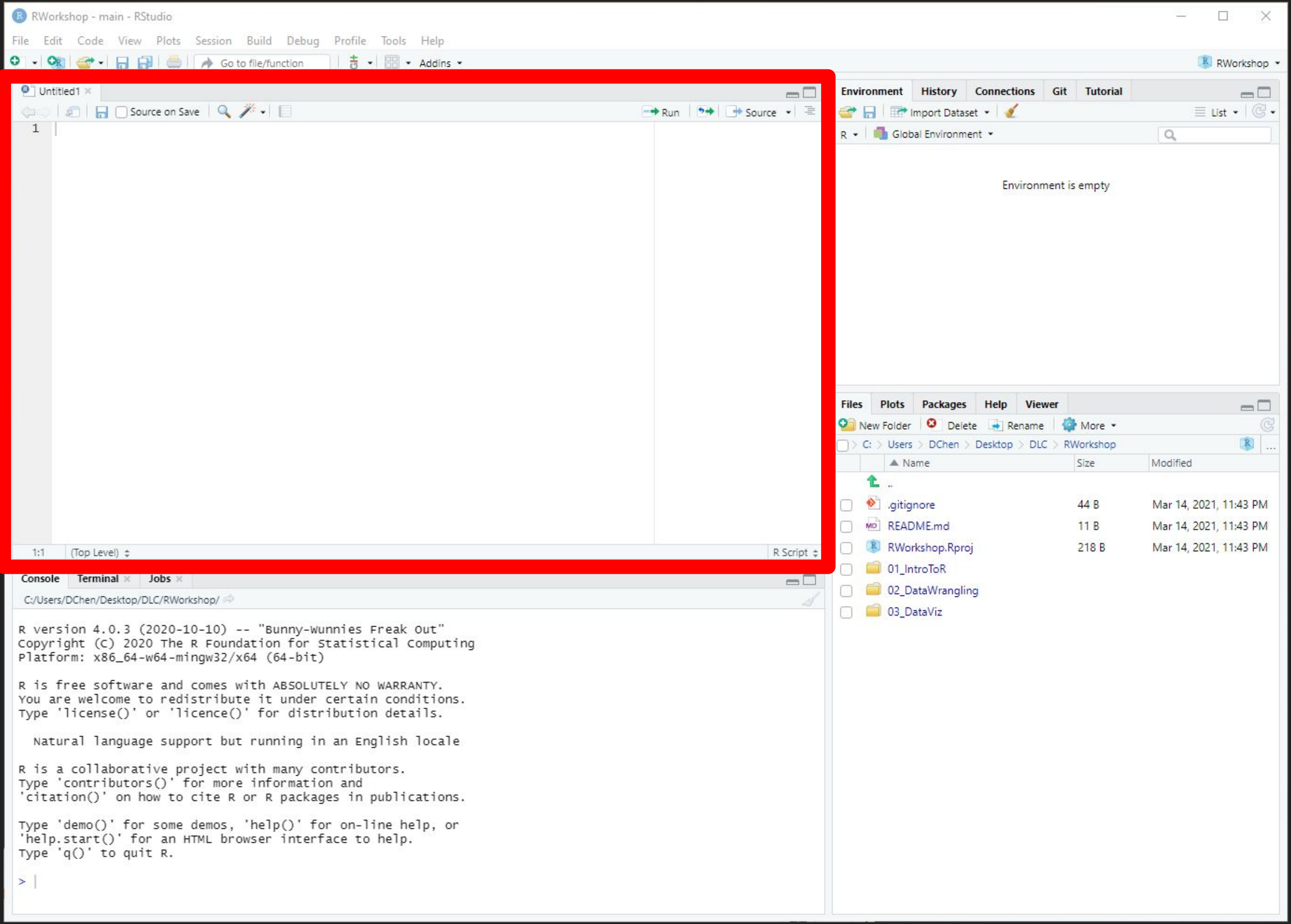
```
R version 4.0.3 (2020-10-10) -- "Bunny-Wunnies Freak Out"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

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.

> |
```

Script Editor

Type code

Run code (output in console)

Saves all code


RWorkshop - main - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
RWorkshop
Untitled1
Source on Save
Run
Source
1:1
(Top Level)
R Script
Environment History Connections Git Tutorial
Import Dataset
List
R
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder
Delete
Rename
More
C: > Users > DChen > Desktop > DLC > RWorkshop
Name
Size
Modified
Console Terminal Jobs
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		

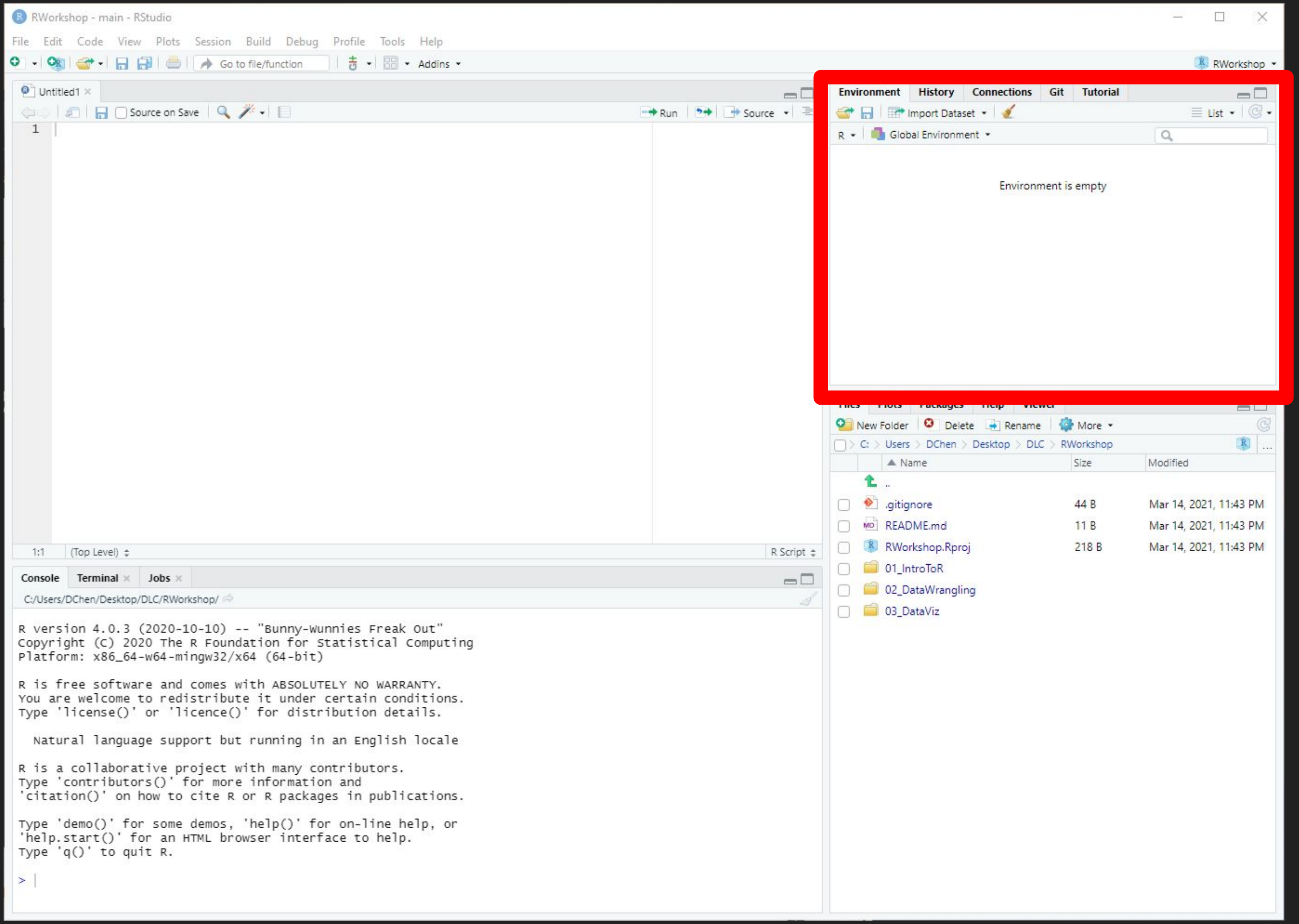
```
R version 4.0.3 (2020-10-10) -- "Bunny-Wunnies Freak Out"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

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.

> 
```

Environment

Import Datasets

Shows all datasets and defined variables

View imported datasets



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		

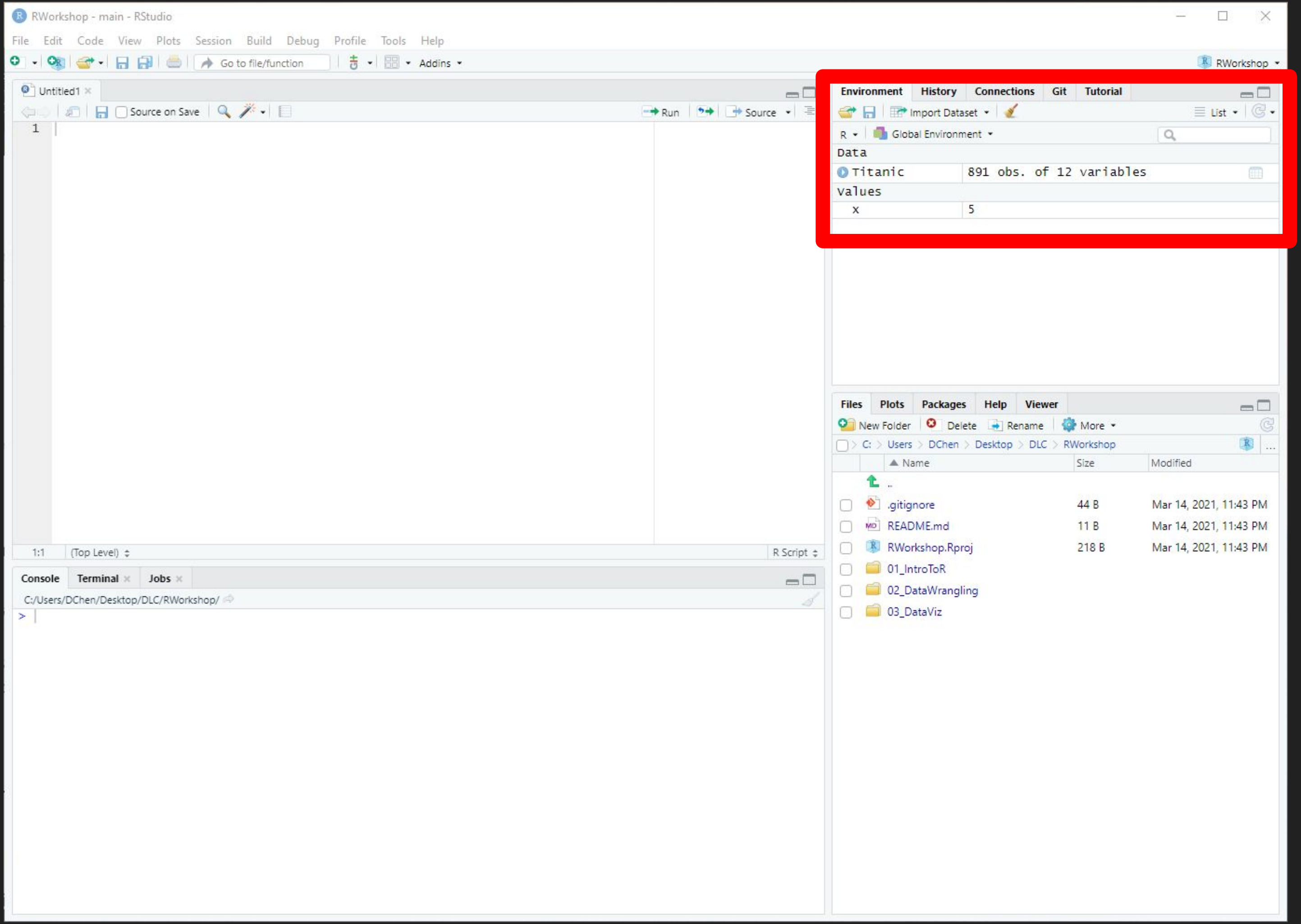
```
C:/Users/DChen/Desktop/DLC/RWorkshop/

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.

> 
```

Environment

Import Datasets

Shows all datasets and defined variables

View imported datasets


RWorkshop - main - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
RWorkshop
Untitled1
Source on Save
Run
Source
Environment History Connections Git Tutorial
Import Dataset
List
Global Environment
Files Plots Packages Help Viewer
New Folder Delete Rename More
Console Terminal Jobs
R Script

```
Data
Titanic    891 obs. of 12 variables
Values
  x        5
```

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		

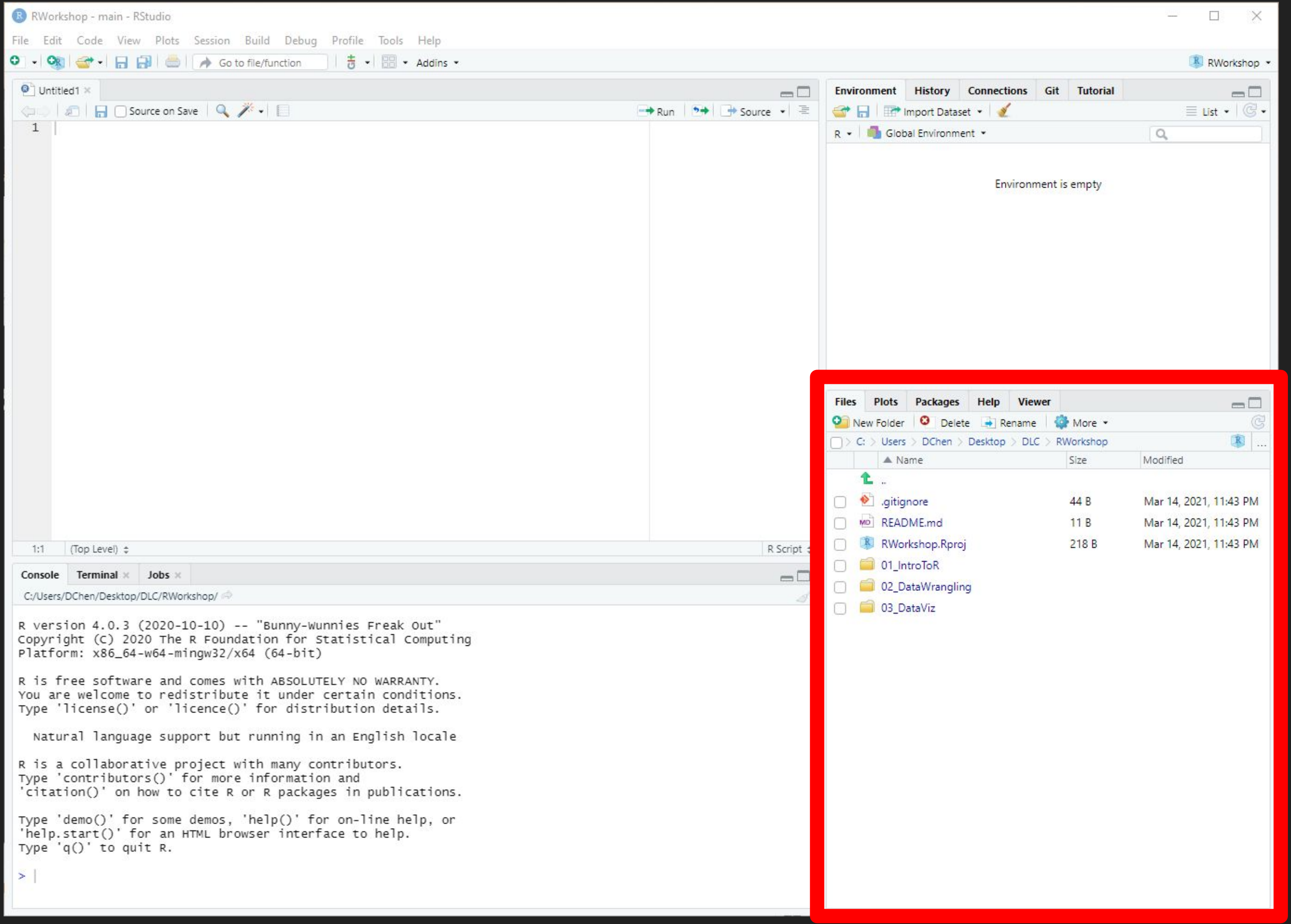
```
C:/Users/DChen/Desktop/DLC/RWorkshop/
>
```

Misc.

Access Files

View created plots

Read help documentation


RWorkshop - main - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
RWorkshop
Untitled1
Source on Save
Run
Source
1:1
(Top Level)
R Script
Environment History Connections Git Tutorial
Import Dataset
List
R
Global Environment
Environment is empty
Console Terminal Jobs
C:/Users/DChen/Desktop/DLC/RWorkshop/
Files Plots Packages Help Viewer
New Folder
Delete
Rename
More
C: Users DChen Desktop DLC RWorkshop
Name
Size
Modified

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		

```
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.

> 
```

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:

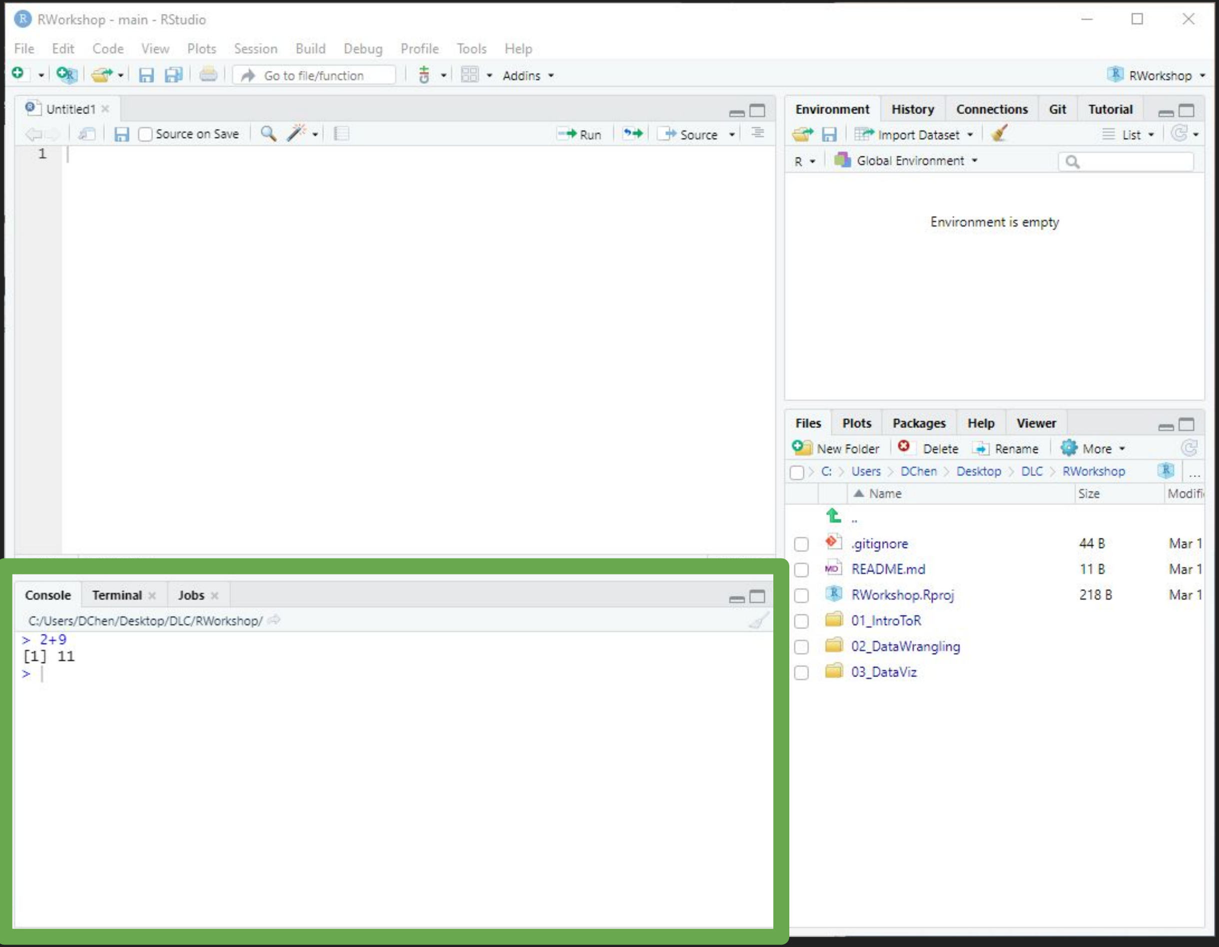
> 2+9

> 15*20

> 20/5

> 2^2

Press `Enter` to run


RWorkshop - main - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
RWorkshop
Untitled1
Source on Save
Run
Source
Environment History Connections Git Tutorial
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: Users DChen Desktop DLC RWorkshop
Name Size Modified
.gitignore 44 B Mar 1
README.md 11 B Mar 1
RWorkshop.Rproj 218 B Mar 1
01_IntroToR
02_DataWrangling
03_DataViz
Console Terminal Jobs
C:/Users/DChen/Desktop/DLC/RWorkshop/

```
> 2+9
[1] 11
>
```

Running Code in the Console

Type the following calculations:

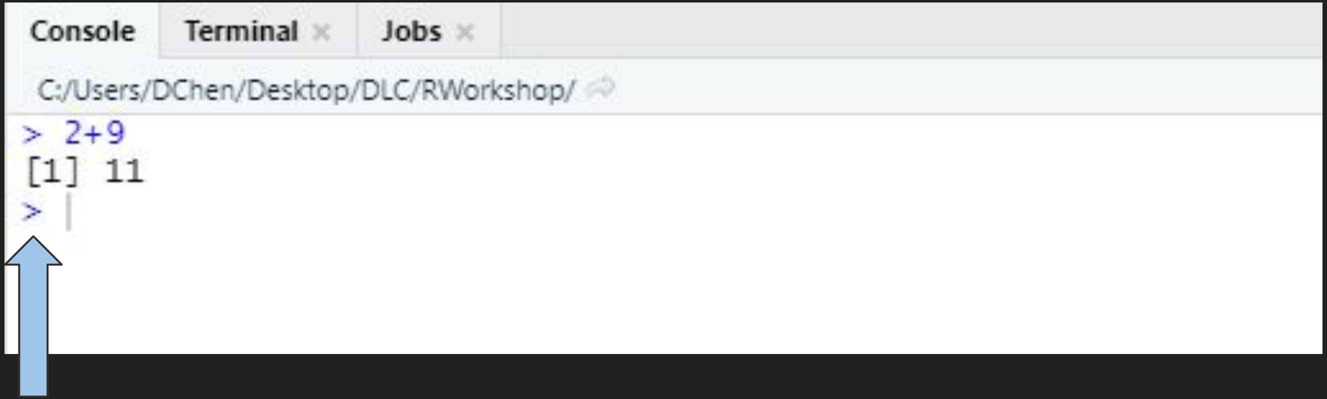
> 2+9

> 15*20

> 20/5

> 2^2

Press `Enter` to run



```
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:

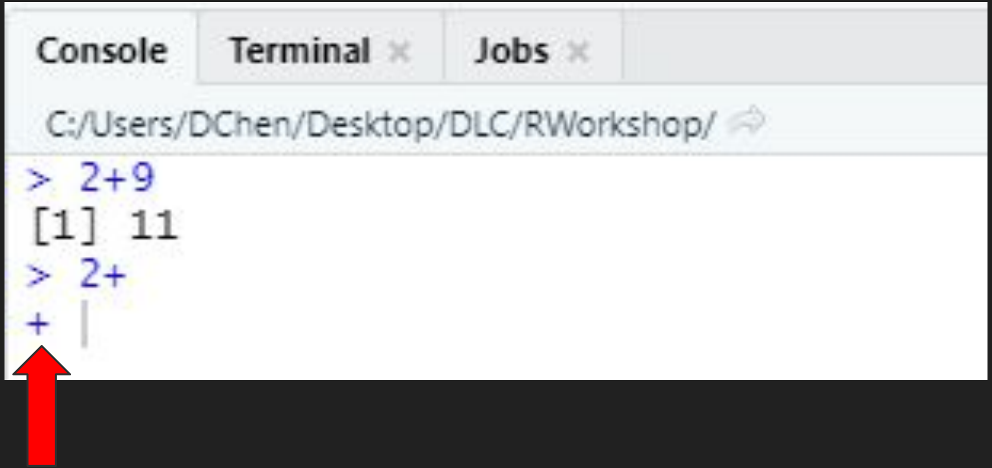
> 2+9

> 15*20

> 20/5

> 2^2

Press `Enter` to run



```
> 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 Console

Type the following calculations:

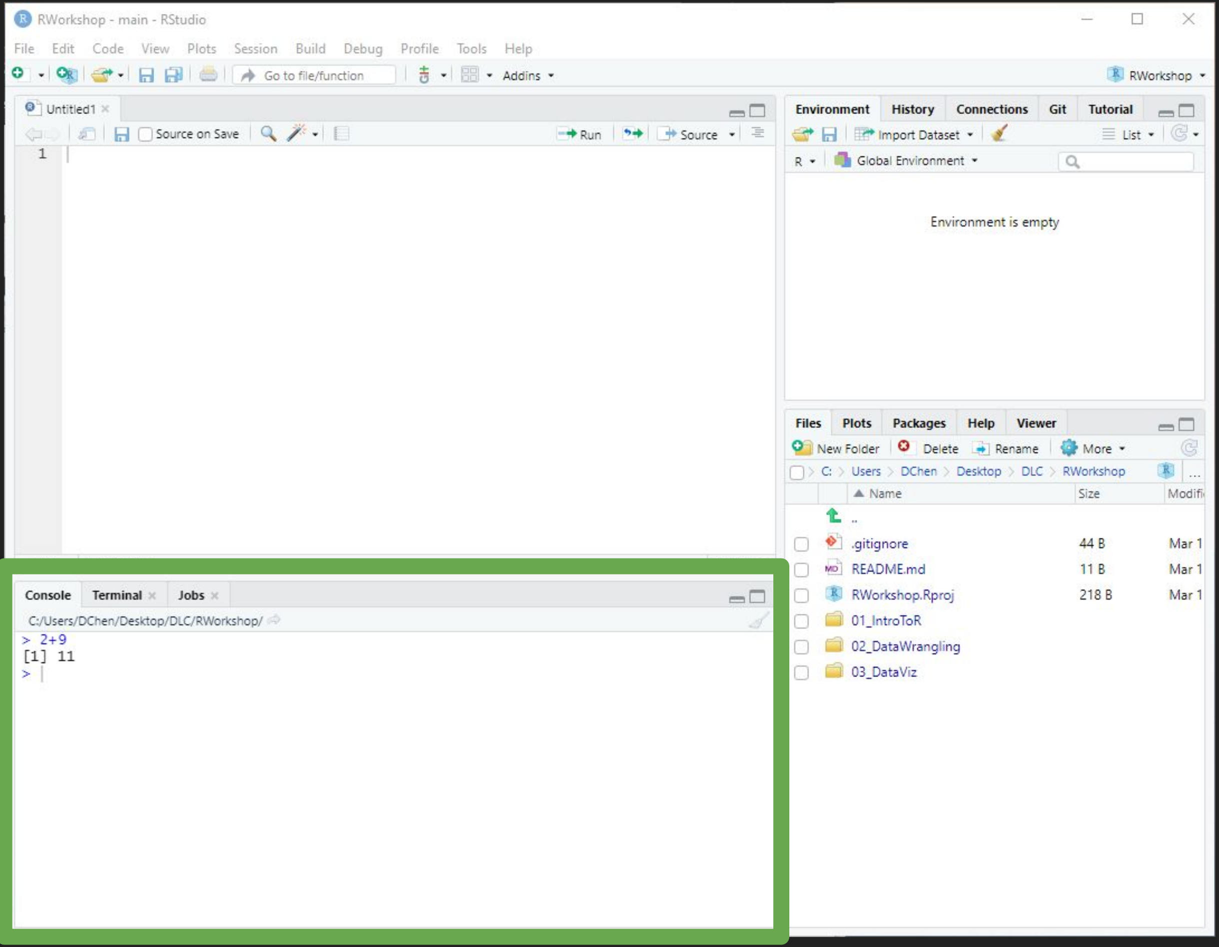
> 2+9

> 15*20

> 20/5

> 2^2

Press `Enter` to run

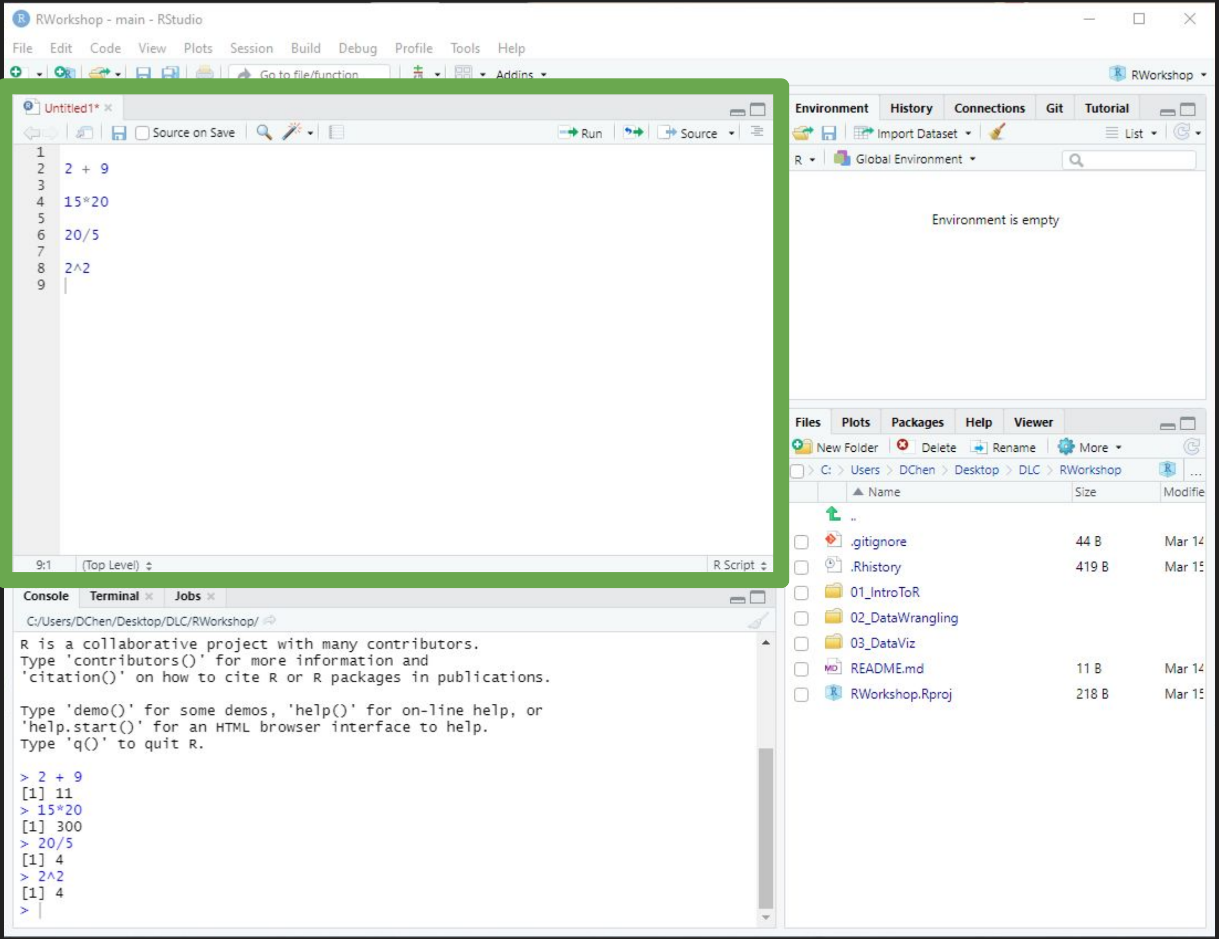

RWorkshop - main - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
RWorkshop
Untitled1
Source on Save
Run
Source
Environment History Connections Git Tutorial
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: > Users > DChen > Desktop > DLC > RWorkshop
Name Size Modified
.gitignore 44 B Mar 1
README.md 11 B Mar 1
RWorkshop.Rproj 218 B Mar 1
01_IntroToR
02_DataWrangling
03_DataViz
Console Terminal Jobs
C:/Users/DChen/Desktop/DLC/RWorkshop/

```
> 2+9
[1] 11
>
```

Running Code in the Script

Code in the script can be saved and ran repeatedly

Output shows up in the console



RWorkshop - main - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Untitled1*
Source on Save
Run
Source
Environment History Connections Git Tutorial
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: Users DChen Desktop DLC RWorkshop
Name Size Modified
.gitignore 44 B Mar 14
.Rhistory 419 B Mar 15
01_IntroToR
02_DataWrangling
03_DataViz
README.md 11 B Mar 14
RWorkshop.Rproj 218 B Mar 15
9:1 (Top Level) R Script
Console Terminal Jobs
C:/Users/DChen/Desktop/DLC/RWorkshop/

```r
2 + 9

15*20

20/5

2^2
```

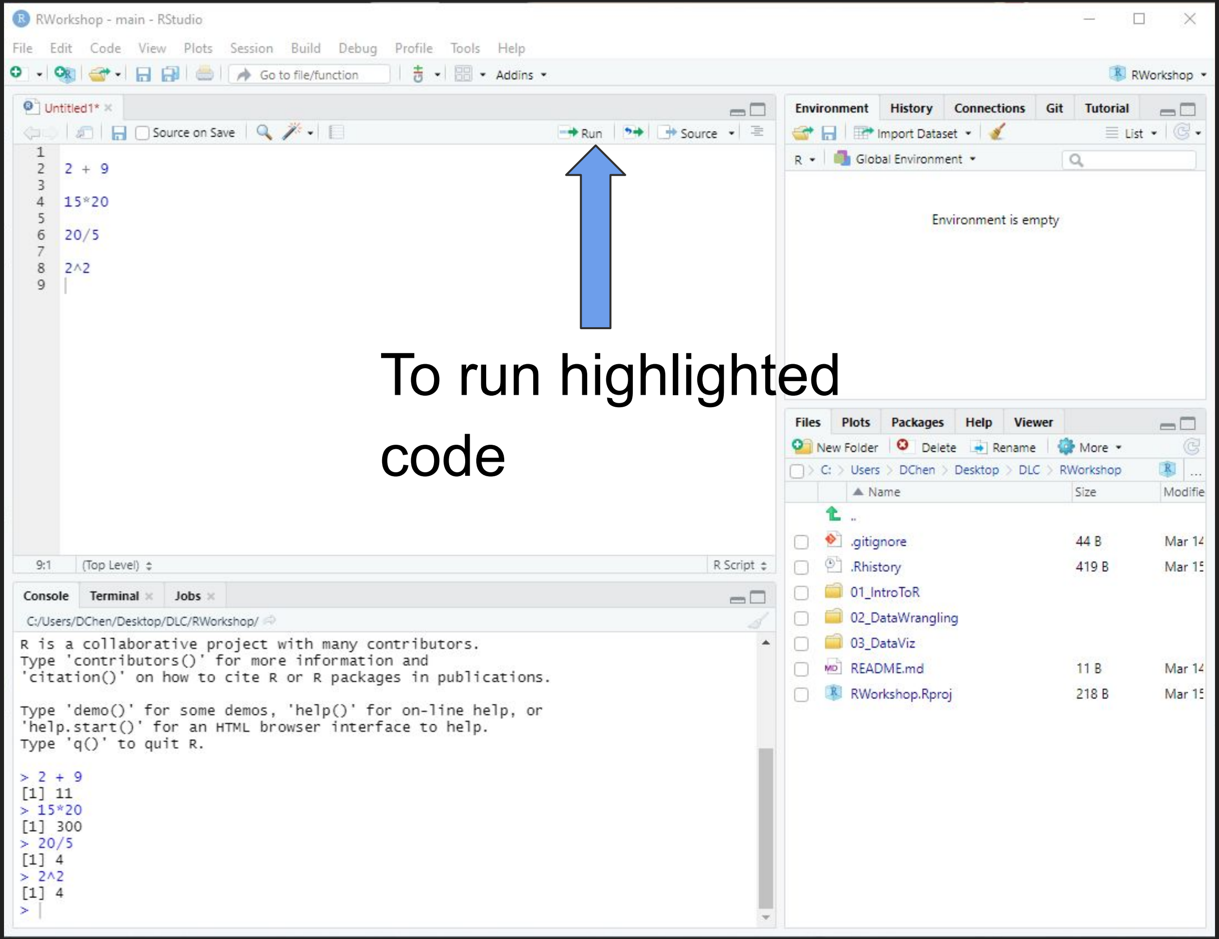
```
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

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.

> 2 + 9
[1] 11
> 15*20
[1] 300
> 20/5
[1] 4
> 2^2
[1] 4
>
```

Running Code in the Script

Type code into the R Script

Click or highlight the line of code and

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


RWorkshop - main - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
RWorkshop
Untitled1*
Source on Save
Run
Source
9:1
(Top Level)
R Script
Environment History Connections Git Tutorial
Import Dataset
List
R
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: > Users > DChen > Desktop > DLC > RWorkshop
Console Terminal Jobs
C:/Users/DChen/Desktop/DLC/RWorkshop/

To run highlighted code

```r
2 + 9

15*20

20/5

2^2
```

```
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.

> 2 + 9
[1] 11
> 15*20
[1] 300
> 20/5
[1] 4
> 2^2
[1] 4
>
```

Name	Size	Modifie
..		
.gitignore	44 B	Mar 14
.Rhistory	419 B	Mar 15
01_IntroToR		
02_DataWrangling		
03_DataViz		
README.md	11 B	Mar 14
RWorkshop.Rproj	218 B	Mar 15

Running Code in the Script

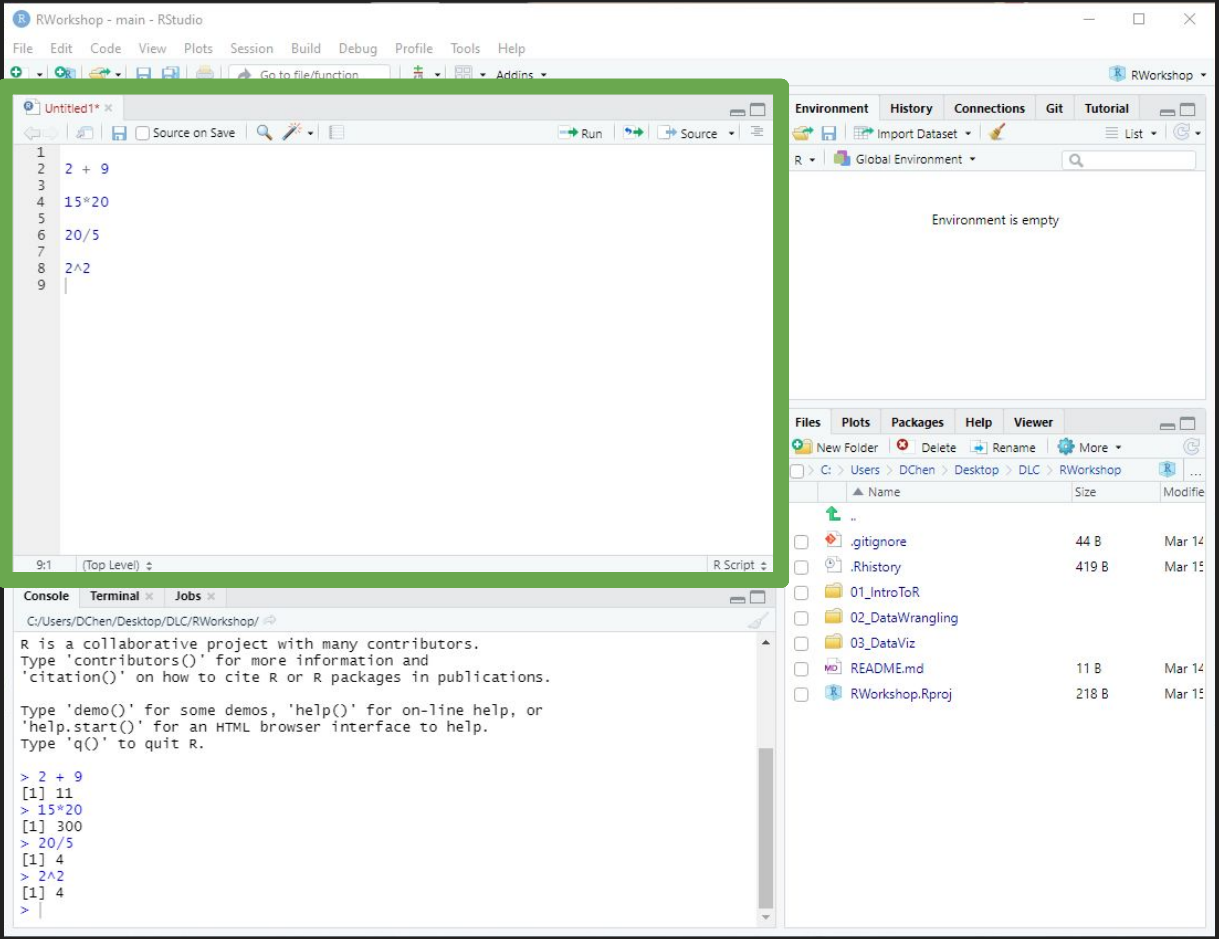
Type the following calculations into the R Script:

> 2+9

> 15*20

> 20/5

> 2^2



RWorkshop - main - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Untitled1*
Source on Save
Run
Source
Environment History Connections Git Tutorial
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: > Users > DChen > Desktop > DLC > RWorkshop
Name Size Modified
.gitignore 44 B Mar 14
.Rhistory 419 B Mar 15
01_IntroToR
02_DataWrangling
03_DataViz
README.md 11 B Mar 14
RWorkshop.Rproj 218 B Mar 15
9:1 (Top Level) R Script
Console Terminal Jobs
C:/Users/DChen/Desktop/DLC/RWorkshop/

```
2 + 9

15*20

20/5

2^2
```

```
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.

> 2 + 9
[1] 11
> 15*20
[1] 300
> 20/5
[1] 4
> 2^2
[1] 4
>
```

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”)
```

Learn more about functions

- Google “<function name> in R”
- Use `help()` or `?` in R to see documentation

```
> help(sqrt)
```

```
> ?sqrt
```

Data Types

Common Data Types

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

(HBC Source)

Common Data Types

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

(HBC Source)

Common Data Types

Data Type	Examples
Numeric	-5, 1, 3.33, 100, pi
Integer	-5L, 1L, 100L
Character	‘words’, “3.33”, ‘TRUE’, “1L”

(HBC Source)

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

(HBC Source)

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

(HBC Source)

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

TRUE

- 5 > 3
- 5 != 3

FALSE

- 5 <= 3
- 8 == 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 - 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!

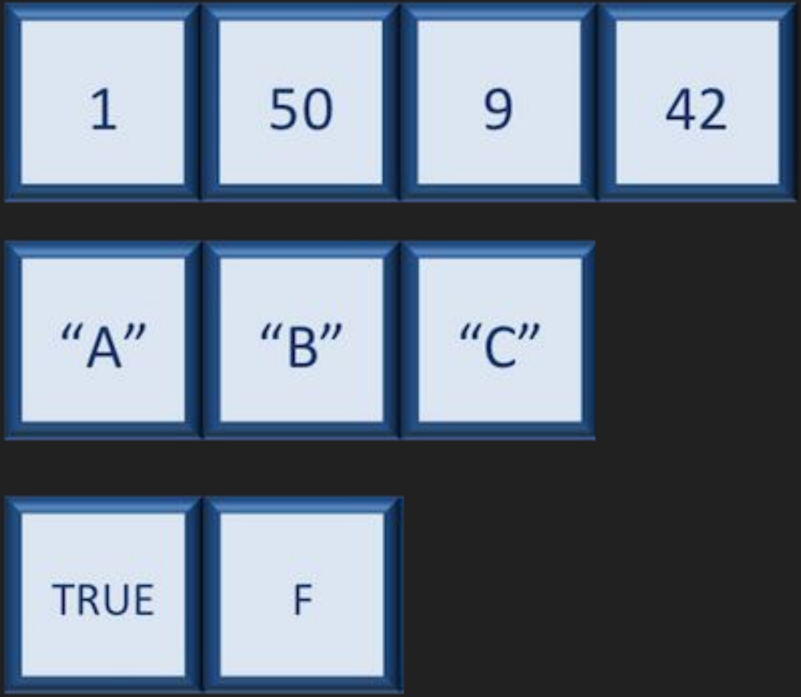
- Hi <- 5

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



(HBC Source)

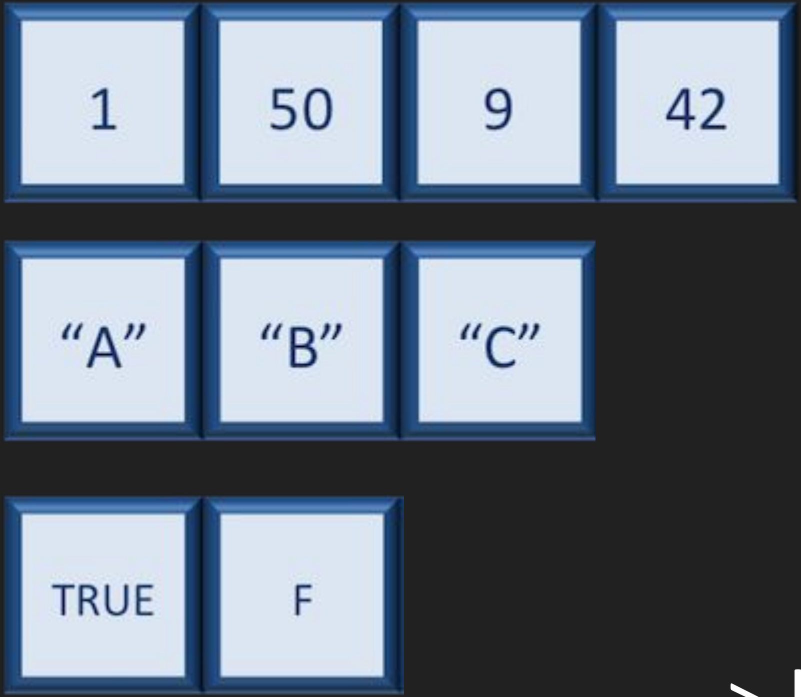
A column in excel

Any length from 1 onwards

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

Defined with function `c()`

Vectors



(HBC Source)

A column in excel

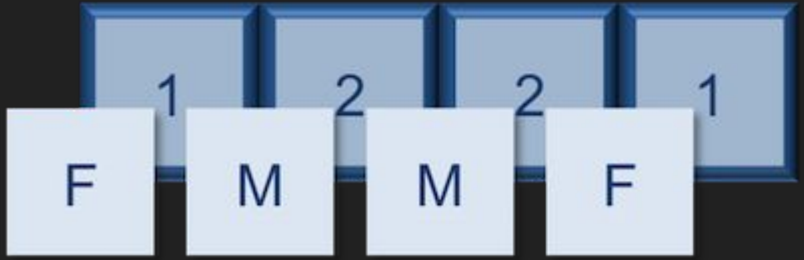
Any length from 1 onwards

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

Defined with function `c()`

```
> lengths <- c(1, 50, 9, 42)
```

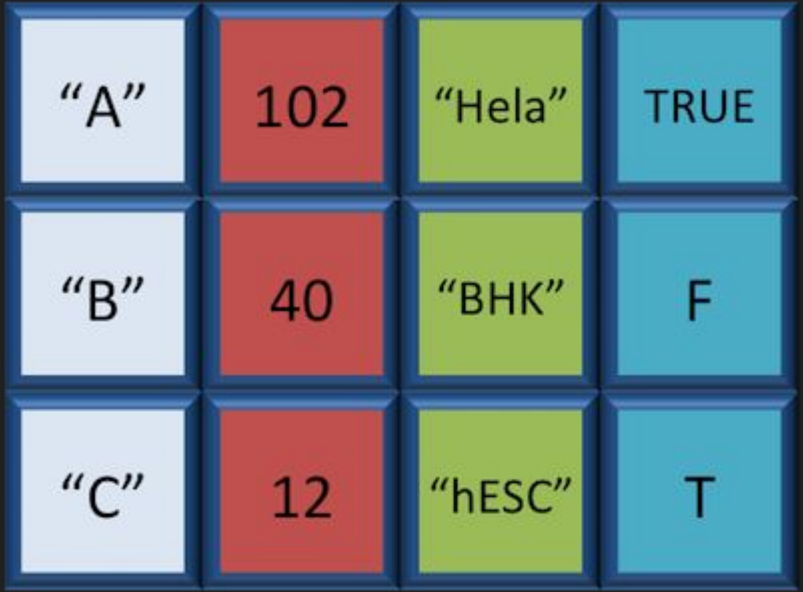
Factors



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

(HBC Source)

Data Frame



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

(HBC Source)

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

Reading in Data

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

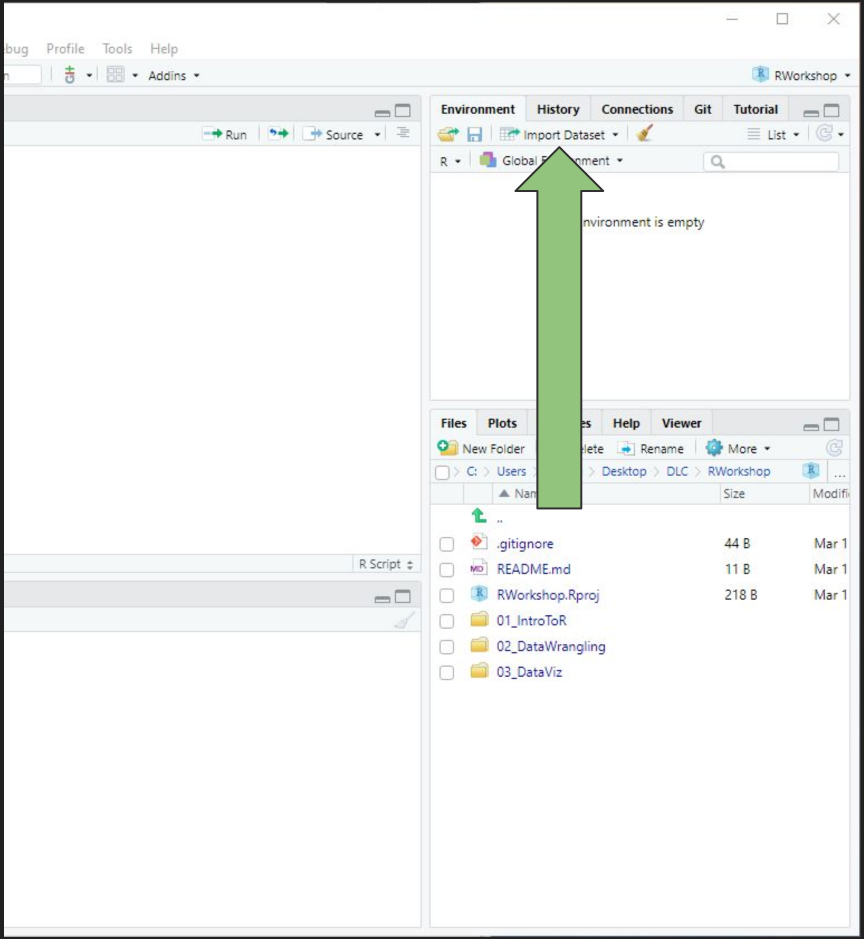
Reading in Datasets

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

(HBC Source)

Reading in Datasets

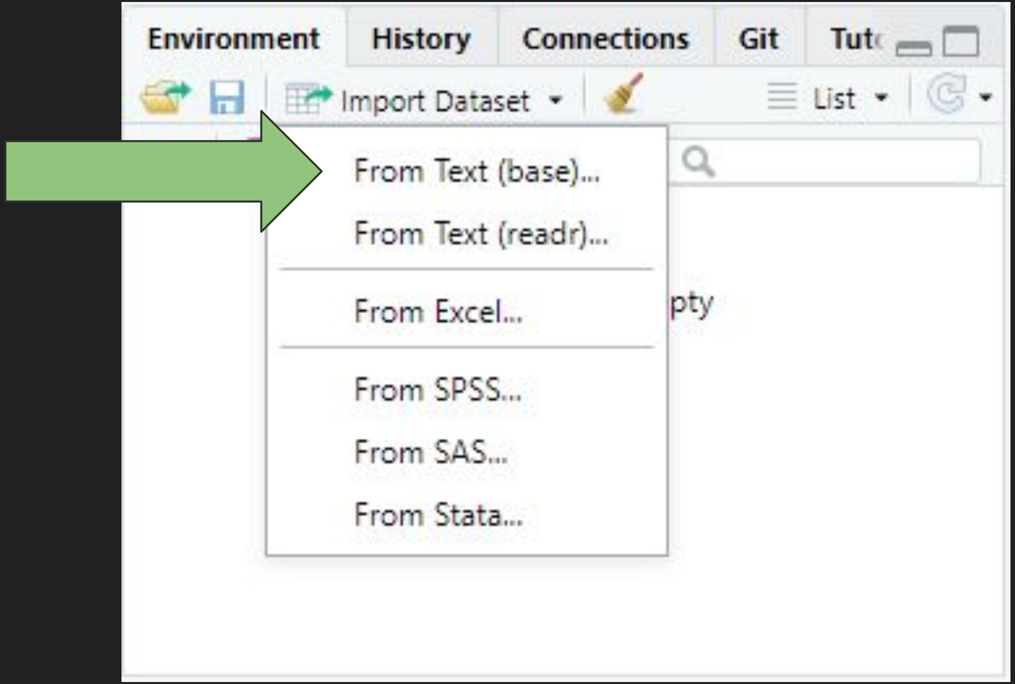
`Import Dataset` in the environment will automatically generate the code for you!



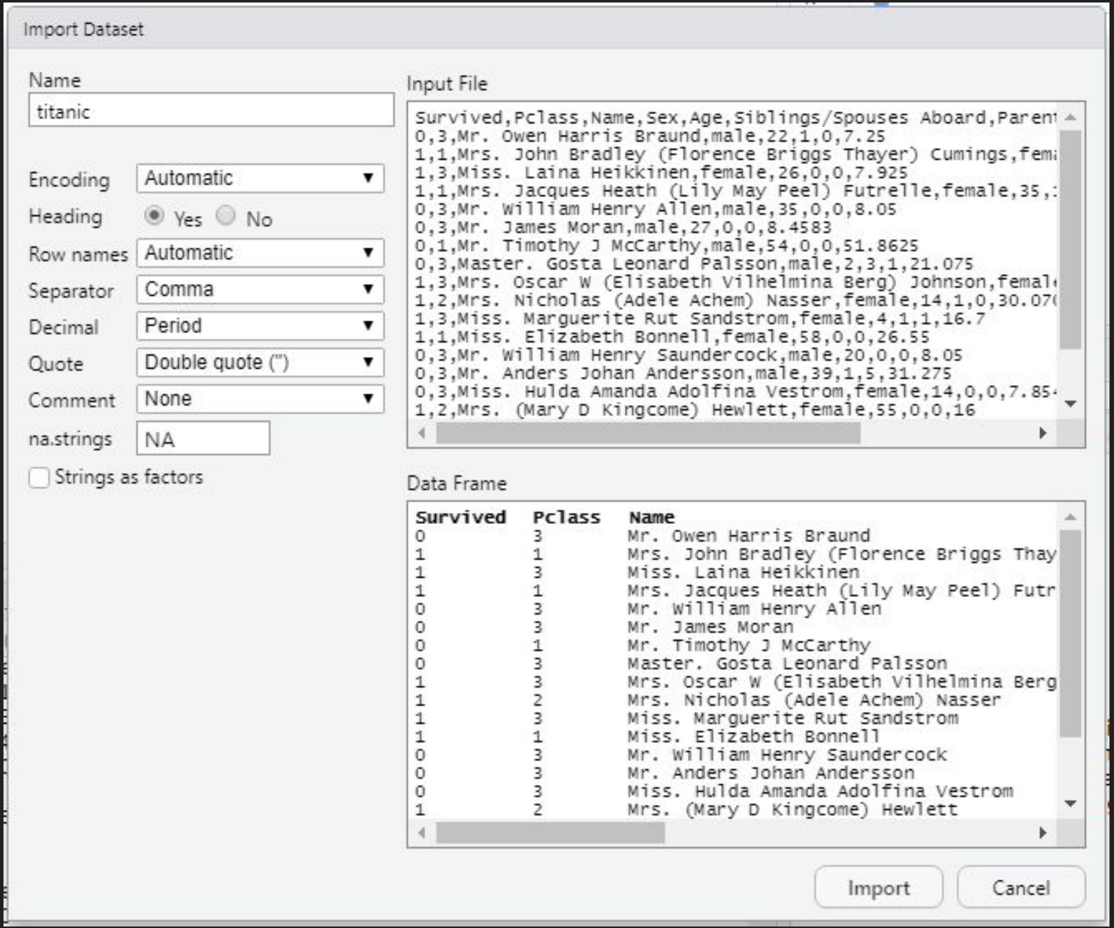
Reading in Datasets

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

Find and select the dataset



Reading in Datasets



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,fema
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 Saundercock,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

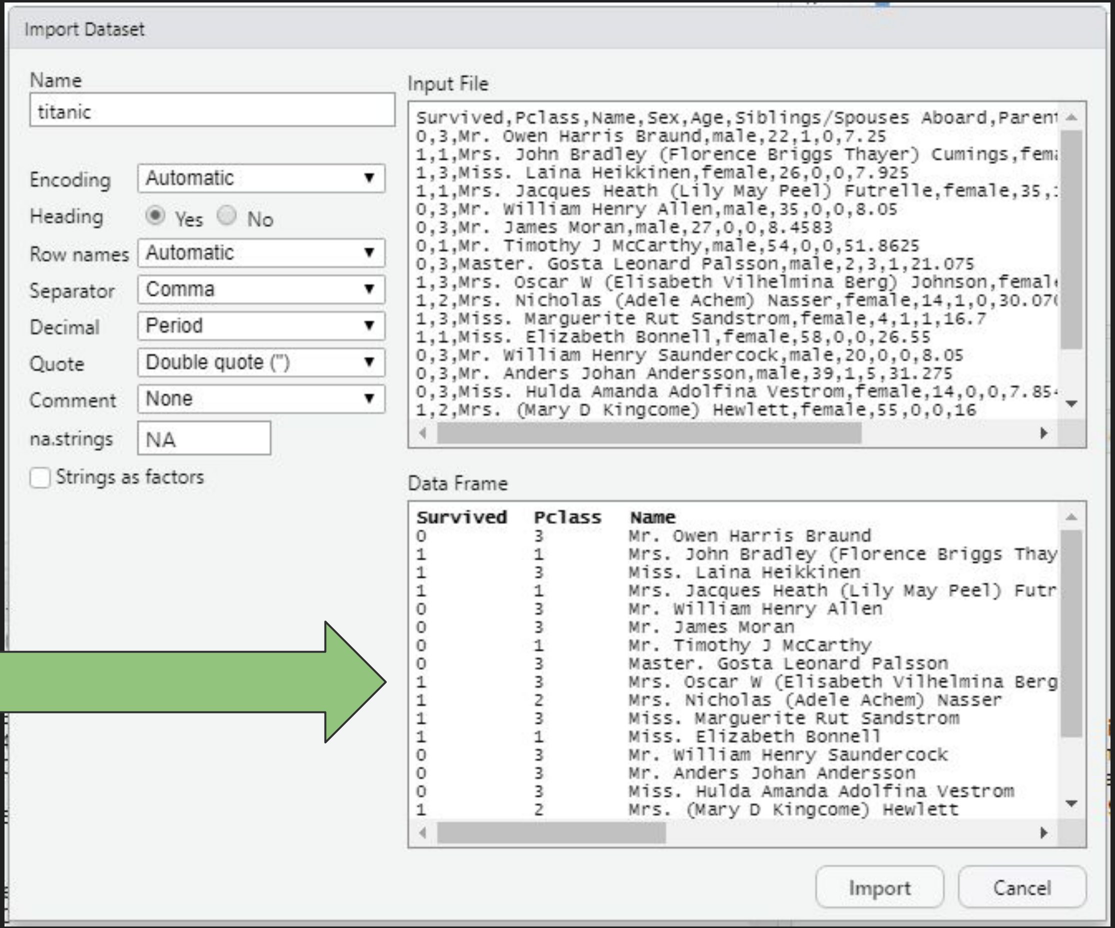
```
Survived   Pclass   Name
0          3        Mr. Owen Harris Braund
1          1        Mrs. John Bradley (Florence Briggs Thay
1          3        Miss. Laina Heikkinen
1          1        Mrs. Jacques Heath (Lily May Peel) Futr
0          3        Mr. William Henry Allen
0          3        Mr.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

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
0,3,Mr. Owen Harris Braund,male,22,1,0,7.25
1,1,Mrs. John Bradley (Florence Briggs Thayer) Cumings,fema
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 Saundercock,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
```

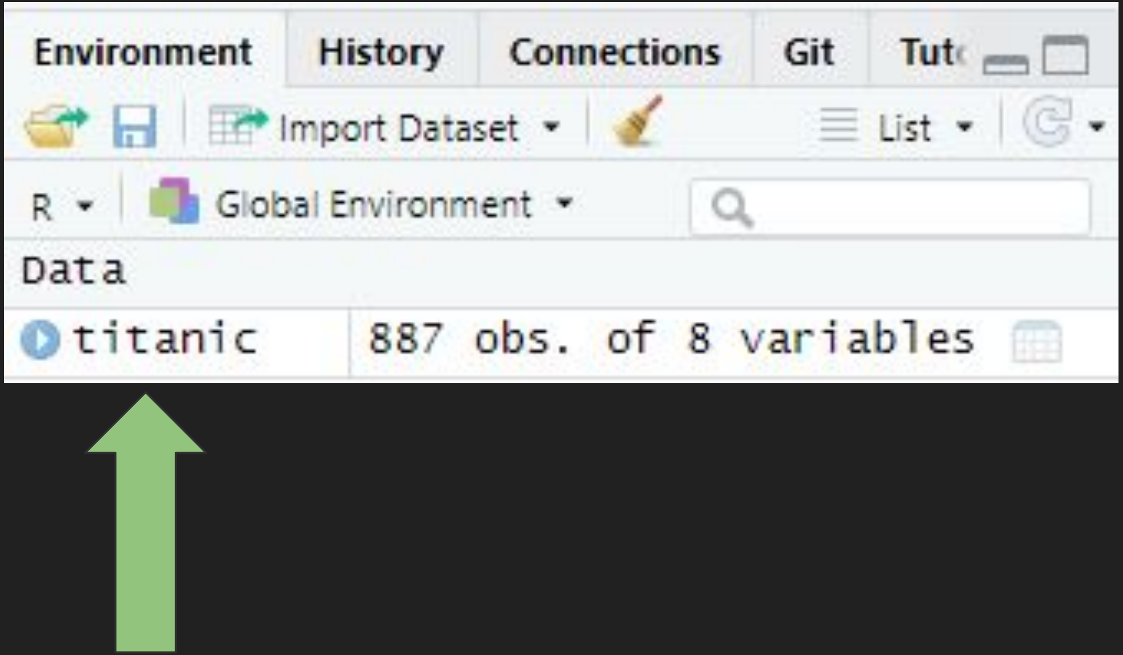
Data Frame

```
Survived    Pclass    Name
0           3         Mr. Owen Harris Braund
1           1         Mrs. John Bradley (Florence Briggs Thay
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
1           2         Mrs. Nicholas (Adele Achem) Nasser
1           3         Miss. Marguerite Rut Sandstrom
1           1         Miss. Elizabeth Bonnell
0           3         Mr. William Henry Saundercock
0           3         Mr. Anders Johan Andersson
0           3         Miss. Hulda Amanda Adolfina Vestrom
1           2         Mrs. (Mary D Kingcome) Hewlett
```

Viewing the Datasets

The data is now in the environment!

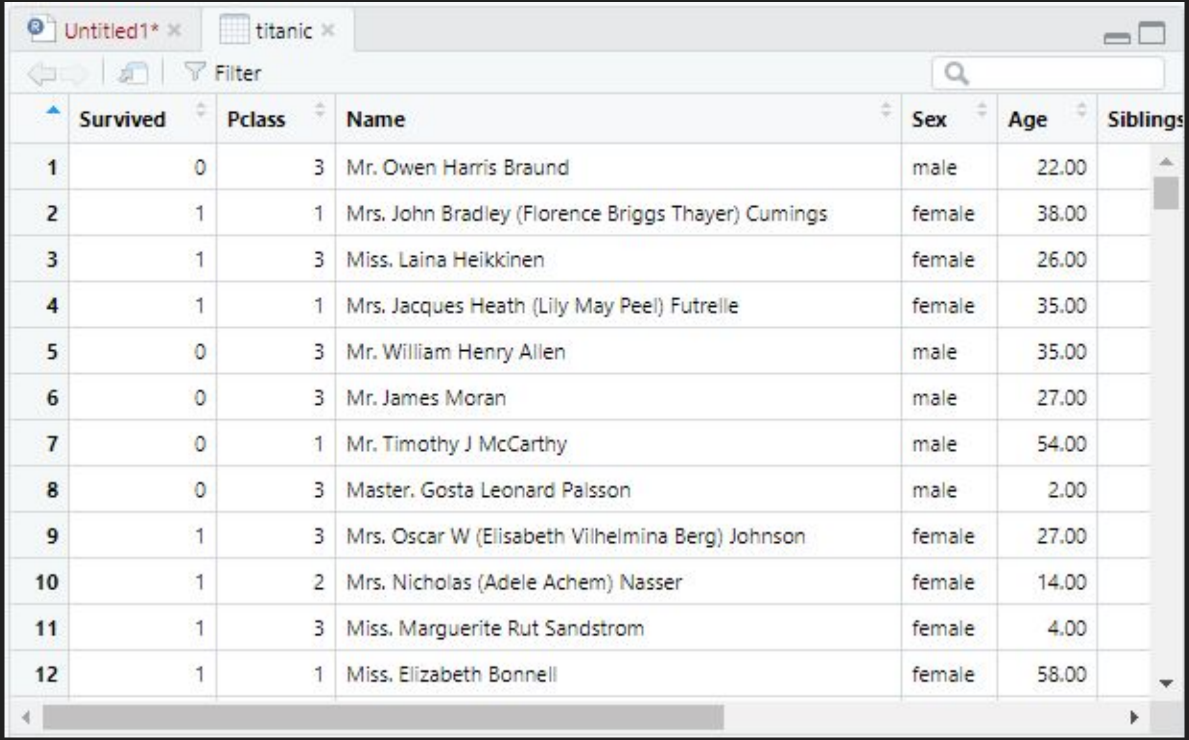
Click the name (titanic) to view the data!



```
titanic    887 obs. of 8 variables
```

Viewing the Datasets

The data view after clicking the name in the environment



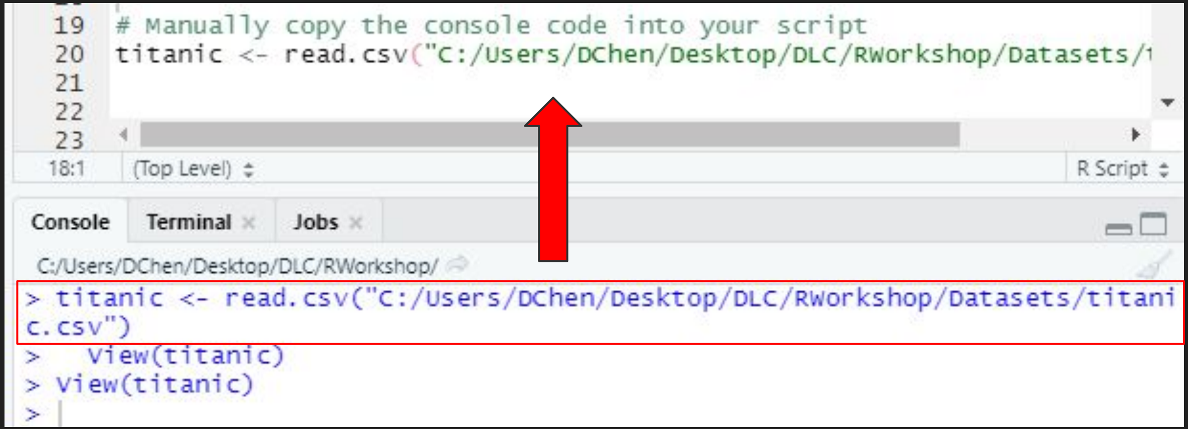
	Survived	Pclass	Name	Sex	Age	Siblings
1	0	3	Mr. Owen Harris Braund	male	22.00	
2	1	1	Mrs. John Bradley (Florence Briggs Thayer) Cumings	female	38.00	
3	1	3	Miss. Laina Heikkinen	female	26.00	
4	1	1	Mrs. Jacques Heath (Lily May Peel) Futrelle	female	35.00	
5	0	3	Mr. William Henry Allen	male	35.00	
6	0	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.



```
# Manually copy the console code into your script
titanic <- read.csv("C:/Users/DChen/Desktop/DLC/RWorkshop/Datasets/t
```

```
> titanic <- read.csv("C:/Users/DChen/Desktop/DLC/RWorkshop/Datasets/titani
c.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)
```

Data Wrangling

Before we start...

Import the `titanic` dataset into R

https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/problem12.html

Install the packages ONCE (to check, skip this step)

```
> install.packages(“tidyverse”)
```

After installing, you must load the packages every time you open R

```
> library(tidyverse)
```

Motivation

What is the purpose of Data Wrangling?

What is the Tidyverse?

Over 87 packages combined together

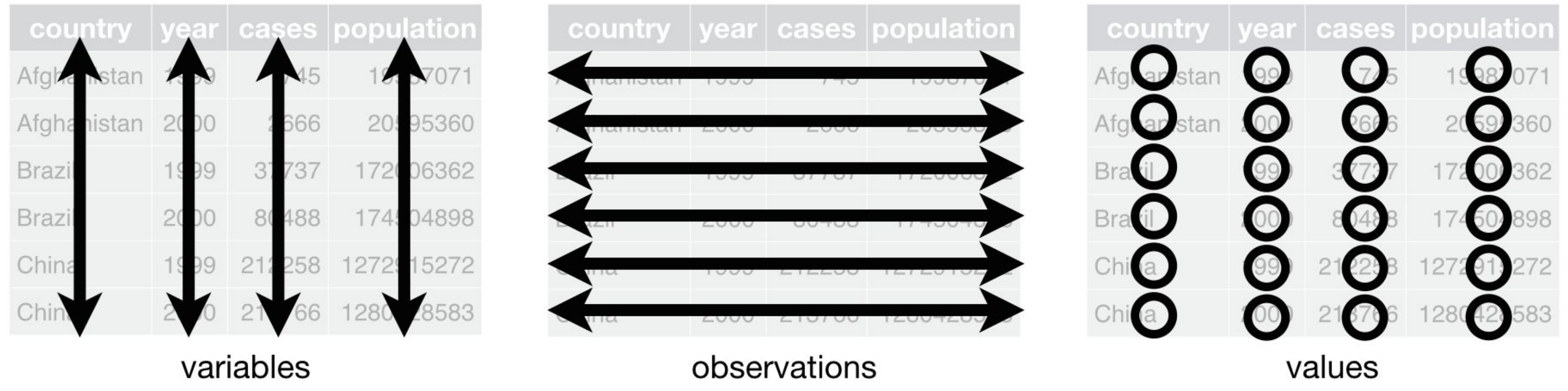
Packages align on data processing philosophy

Encompasses all aspects of a project, including data processing, visualization, modeling

https://cran.r-project.org/web/packages/tidyverse/vignettes/paper.html https://r4ds.had.co.nz/tidy-data.html

Tidyverse Philosophy

“Variables are in columns, observations are in rows, and values are in cells.”



The Fundamental Functions

The Fundamental Functions

select	Select columns

select - Select which variables to keep

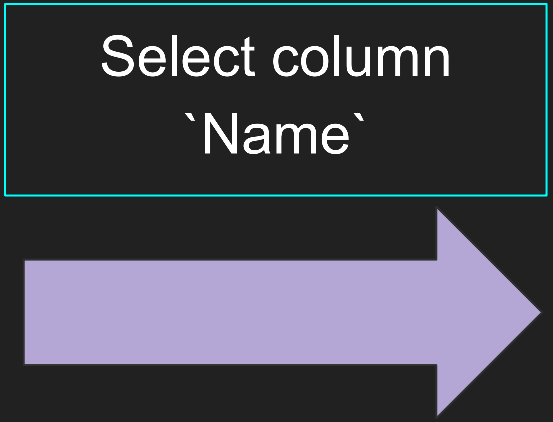
$$\texttt{select(data, }\underbrace{\texttt{`Column1`, `Column2`,...)}}_{\text{Columns to keep or remove (-)}}$$

Use column name, number, or colon

select - Keep columns

select(data, Name)

Name	Major	HW
Matt	math	60
Mary	math	90
Bill	biology	85
Brie	biology	87



Name
Matt
Mary
Bill
Brie

select - Remove Columns

NOT

select(data, -Major)

Name	Major	HW
Matt	math	60
Mary	math	90
Bill	biology	85
Brie	biology	87

Select all columns **except** `Major`

Name	HW
Matt	60
Mary	90
Bill	85
Brie	87

select - Keep Multiple Columns

select(data, Major, HW)

Name	Major	HW
Matt	math	60
Mary	math	90
Bill	biology	85
Brie	biology	87

Select columns `Major` and `HW`

Major	HW
math	60
math	90
biology	85
biology	87

select - Keep Multiple Columns

```
select(data, Major:HW)
```

Name	Major	HW
Matt	math	60
Mary	math	90
Bill	biology	85
Brie	biology	87

Select columns from `Major` to `HW`

$\Rightarrow$

Major	HW
math	60
math	90
biology	85
biology	87

select - Titanic Exercises

1. Select the variables `Survived`, `Sex`, and `Parents.Children.Aboard`

```
> select(titanic, ______, _______, _______)
```

2. Remove the `Name` variable, and store the new results in `Titanic_nameless`

```
> Titanic_nameless <- select(titanic, ________)
```

3. Store the first four columns of `titanic` in a new dataframe. How many different ways can you do this using `select()`?

select - Titanic Exercise Solutions

1. Select the variables `Survived`, `Sex`, and `Parents.Children.Aboard`

```
> select(titanic, Survived, Sex, Parents.Children.Aboard
```

2. Remove the `Name` variable, and store the new results in `Titanic_nameless`

```
> Titanic_nameless <- select(titanic, -Name)
```

3. Store the first four columns of `titanic` in a new dataframe. How many different ways can you do this using `select()`?

select - Titanic Exercise Solutions

3. Store the first four columns of `titanic` in a new dataframe. How many different ways can you do this using `select()`?

```
> t2 <- select(titanic, 1:4)

> t2 <- select(titanic, Survived:Sex)

> t2 <- select(titanic, Survived, Pclass, Name, Sex)

> t2 <- select(titanic, -Age, -Siblings.Aboard, -Parents.Children.Aboard, -Fare)

> t2 <- select(titanic, -c(Age:Fare))
```

The Fundamental Functions

select	Select columns

The Fundamental Functions

select	Select columns
filter	Filter rows

filter - filter the rows

$$\text{filter(data, }\underbrace{\text{<Boolean condition>,...)}}_{\text{Condition on the variables}}$$

filter - filter the rows

$$\texttt{filter(data, }\underbrace{\texttt{<Boolean condition>,...}}_{\text{Condition on the variables}}\texttt{)}$$

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

filter - Keep specific value

```
filter(data, Major == “math”)
```

Name	Major	HW
Matt	math	60
Mary	math	90
Bill	biology	85
Brie	biology	87

Keep rows where Major is “math”

Name	Major	HW
Matt	math	60
Mary	math	90

filter - Keep range of values

```
filter(data, HW <= 85)
```

Name	Major	HW
Matt	math	60
Mary	math	90
Bill	biology	85
Brie	biology	87

Keep rows where `HW` is less than or equal to 85

Name	Major	HW
Matt	math	60
Bill	biology	85

filter - Multiple Conditions

filter(data, HW <= 85, Major == “math”)

Name	Major	HW
Matt	math	60
Mary	math	90
Bill	biology	85
Brie	biology	87

Keep rows where `HW` is less than or equal to 85, **AND** `Major` is math

Name	Major	HW
Matt	math	60

filter - Multiple Conditions

AND

```
filter(data, HW <= 85 & Major == “math”)
```

Name	Major	HW
Matt	math	60
Mary	math	90
Bill	biology	85
Brie	biology	87

Keep rows where `HW` is less than or equal to 85, **AND** `Major` is math

Name	Major	HW
Matt	math	60

filter - Multiple Conditions

OR

```
filter(data, HW <= 85 | Major == “biology”)
```

Name	Major	HW
Matt	math	60
Mary	math	90
Bill	biology	85
Brie	biology	87

Keep rows where `HW` is less than or equal to 85, **OR** `Major` is biology

Name	Major	HW
Matt	math	60
Bill	biology	85
Brie	biology	87

filter - Titanic Exercises

1. Create a new dataset with only those who survived (`Survived` value of 1)

```
> survived <- filter(titanic, ______ == _____)
```

2. Only show rows with 1st class (`Pclass`) females (`Sex`)

```
> filter(titanic, _____ == 1 _ Sex == _____)
```

3. How many rows are there if you keep only males between ages 20-30?

filter - Titanic Exercise Solutions

1. Create a new dataset with only those who survived (`Survived` value of 1)

```
> survived <- filter(titanic, Survived == 1)
```

```
> survived <- filter(titanic, Survived == “1”)
```

filter - Titanic Exercise Solutions

1. Create a new dataset with only those who survived (`Survived` value of 1)

```
> survived <- filter(titanic, Survived == 1)
```

```
> survived <- filter(titanic, Survived == “1”)
```

2. Only show rows with 1st class (`Pclass`) females (`Sex`)

```
> filter(titanic, Pclass == 1 & Sex == “female”)
```

```
> filter(titanic, Pclass == 1, Sex == “female”)
```

filter - Titanic Exercise Solutions

3. How many rows are there if you keep only males between ages 20-30?

```
> tm <- filter(titanic, Sex == “male”, Age >= 20, Age <= 30)

> nrow(tm) # 223

> t2 <- filter(titanic, Sex == “male”)

> t3 <- filter(t2, Age >= 20)

> t4 <- filter(t3, Age <= 30)
```

The Fundamental Functions

select	Select columns
filter	Filter rows

Pipes

%>%

Pipes %>%

Pipes (%>%) insert the data into the next line

```
function(data)  ==  data %>%
                      function(.)
```

Pipes %>%

```
sqrt(2)          ==    2 %>% sqrt()
```

```
sqrt(sqrt(2))    ==    2 %>%
                         sqrt() %>%
                         sqrt()
```

Pipes %>%

Without pipes, multi-line processing is hard to read

```
df1 <- select(df, A, B, C)

df2 <- filter(df1, A == 1, B > 2)

df3 <- filter(df2, C == 1 | C == 5)
```

Creates unnecessary intermediate datasets

Pipes %>%

Without pipes, multi-line processing is hard to read

```
df1 <- select(df, A, B, C)

df2 <- filter(df1, A == 1, B > 2)

df3 <- filter(df2, C == 1 | C == 5)
```

```
df1 <- filter(filter(select(df, A, B, C), A ==
1, B > 2), C == 1 | C == 5)
```

Hard to read

Pipes %>%

With pipes, one line is one logical step

`df1 <- df %>%`	Start with dataset `df`

Pipes %>%

With pipes, one line is one logical step

`df1 <- df %>%`	Start with dataset `df`
`    select(A, B, C) %>%`	Select variables A, B, C

Pipes %>%

With pipes, one line is one logical step

`df1 <- df %>%`	Start with dataset `df`
`select(A, B, C) %>%`	Select variables A, B, C
`filter(A == 1, B > 2) %>%`	Keep rows with A == 1 and B > 2

Pipes %>%

With pipes, one line is one logical step

```
df1 <- df %>%
    select(A, B, C) %>%
    filter(A == 1, B > 2) %>%
    filter(C == 1 | C == 5)
```

Start with dataset `df`

    Select variables A, B, C

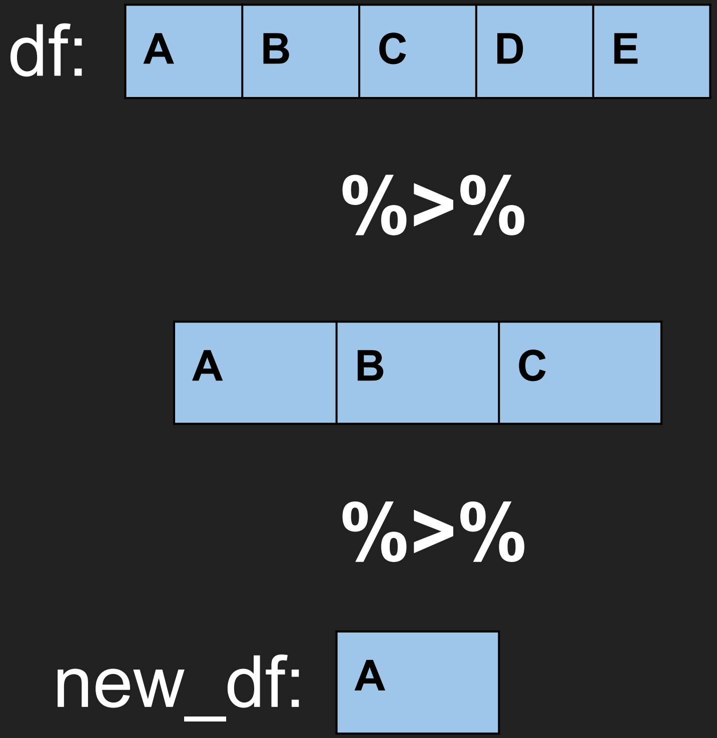
    Keep rows with A == 1 and B > 2

    Keep rows with C == 1 or C == 5

(Store results in df1)

Pipes %>%

```
new_df <- df %>%
    select(A, B, C) %>%
    select(A)
```



Pipe %>% - Titanic Exercise

1. Using pipes, store the first four columns and keep only 1st class (`Pclass`)

```
> firstclass <- _________ %>%
>   select(1_4) _____
>   filter(Pclass __ ______)
```

Pipe %>% - Titanic Exercise Solution

1. Using pipes, store the first four columns and keep only 1st class (`Pclass`)

```
> firstclass <- titanic %>%
>   select(1:4) %>%
>   filter(Pclass == 1)
```

The Fundamental Functions

select	Select columns
filter	Filter rows

The Fundamental Functions

select	Select columns
filter	Filter rows
mutate	Create new variables

mutate - Create new variables/columns

$$\texttt{mutate(., }\underbrace{\texttt{new\_variable = value, ...}}_{\boxed{\substack{\text{Computation of new variable}\\ \text{Can be based on another variable,}\\ \text{a function, a single value, etc.}}}}\texttt{)}$$

mutate - Create var with constant value

```
data %>% mutate(Sect = 1)
```

Name	HW	Ex
Matt	60	75
Mary	90	100
Bill	85	60
Brie	87	85

Create new variable `Sect` where all values are 1

Name	HW	Ex	Sect
Matt	60	75	1
Mary	90	100	1
Bill	85	60	1
Brie	87	85	1

mutate - Create var based on another var

```
data %>% mutate(Ex2 = Ex/100)
```

Name	HW	Ex
Matt	60	75
Mary	90	100
Bill	85	60
Brie	87	85

Create new variable `Ex2` as `Ex` divided by 100

Name	HW	Ex	Ex2
Matt	60	75	0.75
Mary	90	100	1.00
Bill	85	60	0.60
Brie	87	85	0.85

mutate - Create var based on other variables

```
data %>% mutate(Final = (HW + Ex)/2)
```

Name	HW	Ex
Matt	60	75
Mary	90	100
Bill	85	60
Brie	87	85

Create new variable \`Final\` as average of \`HW\` and \`Ex\`

Name	HW	Ex	Final
Matt	60	75	67.5
Mary	90	100	95
Bill	85	60	72.5
Brie	87	85	86

mutate - Titanic Exercises

1. `Fare` is currently in pounds. Convert it to dollars (1 pound = 1.37 dollars)

```
> titanic %>% mutate(FareDollars = _____*______)
```

2. How many family members total (combine `Siblings.Spouses.Aboard` and `Parents.Children.Aboard`) were on board for each passenger?

```
> titanic %>% mutate(titanic, Family = _____ + ______)
```

3. Using pipes: remove the name, keep only 1st class survivors, subtract 20 from `Age`. Store these results in a dataset named `AgeAdjusted`.

mutate - Titanic Exercise Solutions

1. `Fare` is currently in pounds. Convert it to dollars (1 pound = 1.37 dollars)

```
> titanic %>% mutate(FareDollars = Fare*1.37)
```

2. How many family members total (combine `Siblings.Spouses.Aboard` and `Parents.Children.Aboard`) were on board for each passenger?

```
> titanic %>% mutate(titanic, Family =
Siblings.Spouses.Aboard + Parents.Children.Aboard)
```

mutate - Titanic Exercise Solutions

3. Using pipes: remove the name, keep only 1st class survivors, subtract 20 from `Age`. Store these results in a dataset named `AgeAdjusted`.

```
> AgeAdjusted <- titanic %>%
>   select(-Name) %>%
>   filter(Pclass == 1, Survived == 1) %>%
>   mutate(Age = Age - 20)
```

The Fundamental Functions

select	Select columns
filter	Filter rows
mutate	Create new variables

The Fundamental Functions

select	Select columns
filter	Filter rows
mutate	Create new variables
summarize	Compute summary variable

summarize - Create new aggregated variables

$$\text{summarize(., }\underbrace{\text{new\_variable = function(\_), ...)}}_{\boxed{\begin{array}{c}\text{Computation of new variable}\\ \text{Function that returns one value}\\ \text{(e.g. mean(), median(), max())}\end{array}}}$$

summarize - Create new aggregated variables

summarize() computes a single value or column

mutate() adds a variable to existing dataset

summarize - Compute mean of dataset

```
data %>% summarize(HW_mean = mean(HW))
```

Name	HW
Matt	60
Mary	90
Bill	85
Brie	87

Calculate mean HW score

HW_mean
80.5

summarize - Count total number of observations

```
data %>% summarize(Count = n())
```

Name	HW
Matt	60
Mary	90
Bill	85
Brie	87

Calculate number of observations

Count
4

The Fundamental Functions

select	Select columns
filter	Filter rows
mutate	Create new variables
summarize	Compute summary variable

The Fundamental Functions

select	Select columns
filter	Filter rows
mutate	Create new variables
summarize	Compute summary variable
group_by	Split data by groups

group_by - split data by groups

$$\text{group\_by(., }\underbrace{\text{`id\_var`, ...)}}_{\boxed{\text{Split dataset by `id\_var` (or multiple variables)}}}$$

Dataset split based on unique values of `id_var`

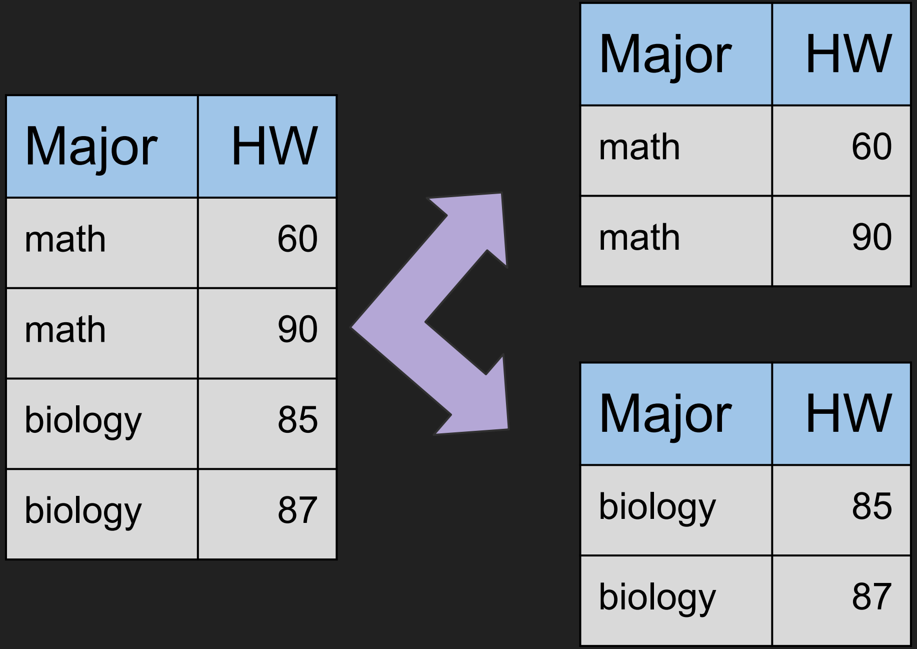
group_by - split data by groups

```
data %>% group_by(., Major)
```

Major	HW
math	60
math	90
biology	85
biology	87

group_by - split data by groups

```
data %>% group_by(., Major)
```



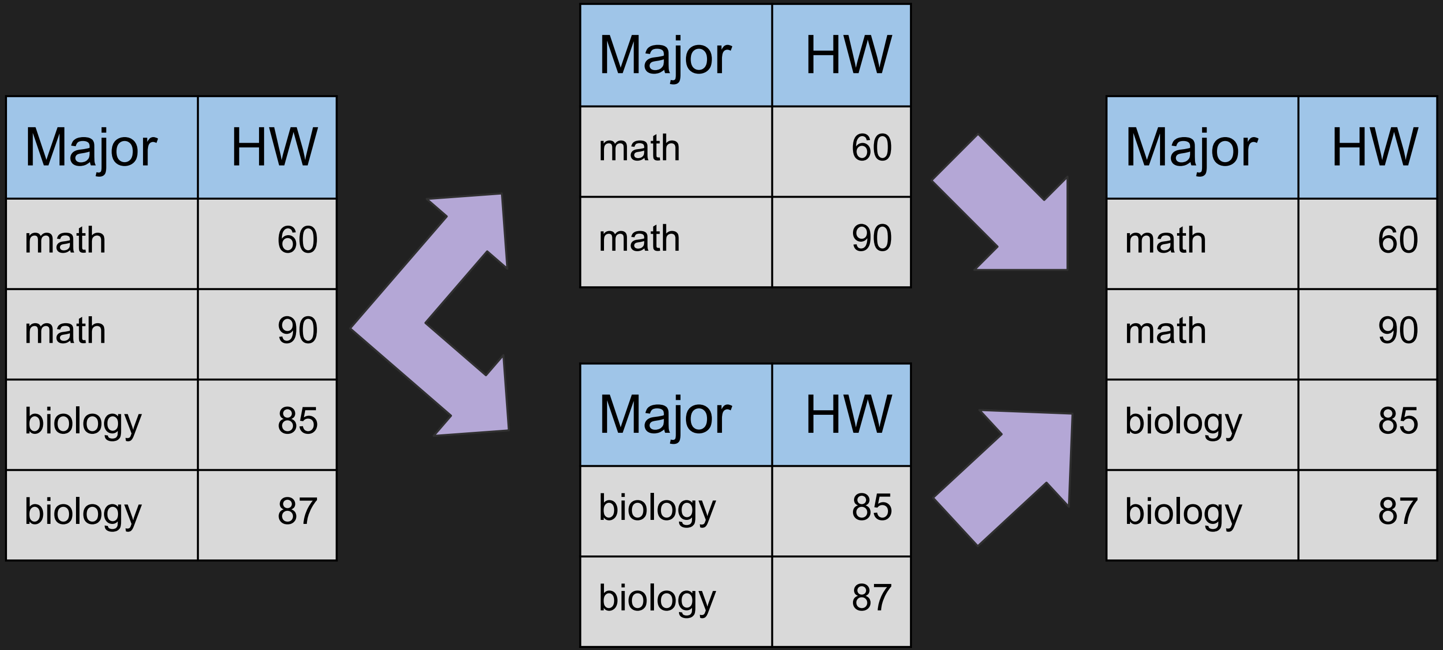
Major	HW
math	60
math	90
biology	85
biology	87

Major	HW
math	60
math	90

Major	HW
biology	85
biology	87

group_by - split data by groups

```
data %>% group_by(., Major)
```



Major	HW
math	60
math	90
biology	85
biology	87

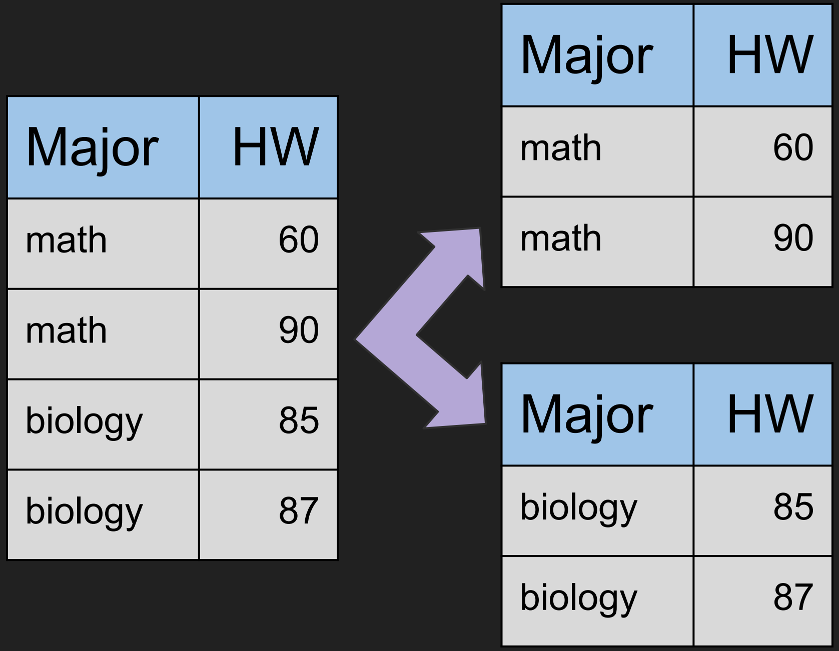
Major	HW
math	60
math	90

Major	HW
biology	85
biology	87

Major	HW
math	60
math	90
biology	85
biology	87

group_by() - split data by groups

```
group_by(., Major)
```



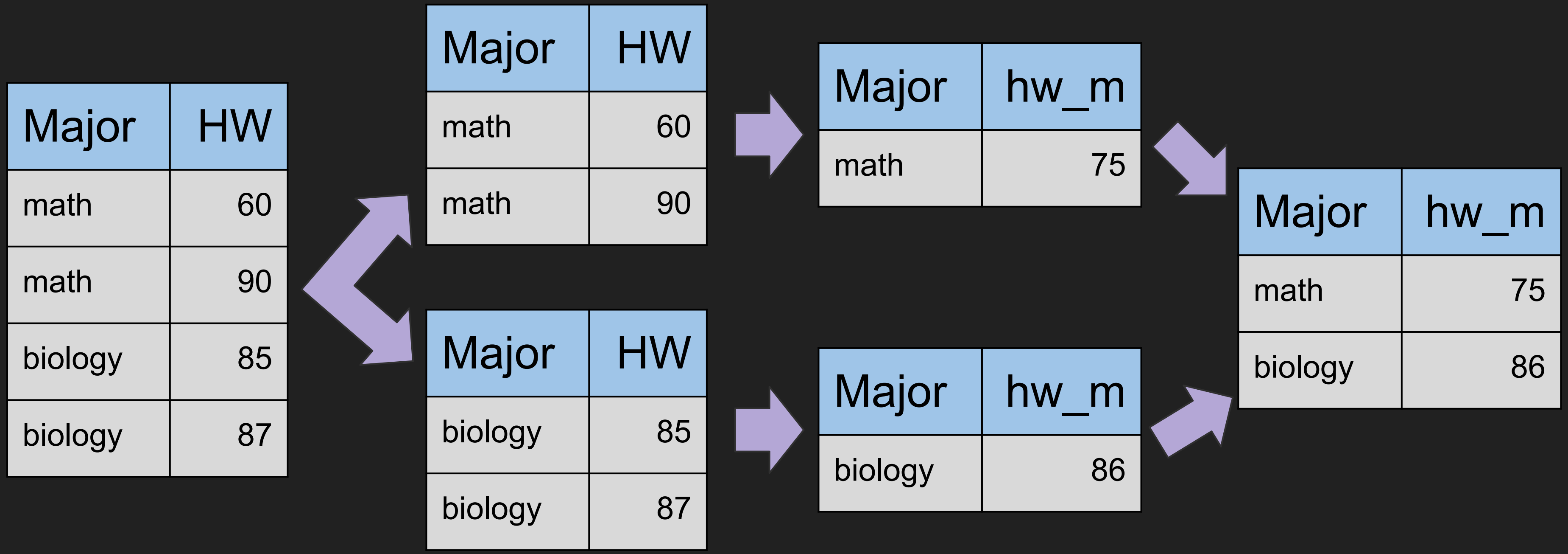
Major	HW
math	60
math	90
biology	85
biology	87

Major	HW
math	60
math	90

Major	HW
biology	85
biology	87

group_by() - split data by groups

```
group_by(., Major) %>% summarize(hw_m = mean(HW))
```



Major	HW
math	60
math	90
biology	85
biology	87

Major	HW
math	60
math	90

Major	HW
biology	85
biology	87

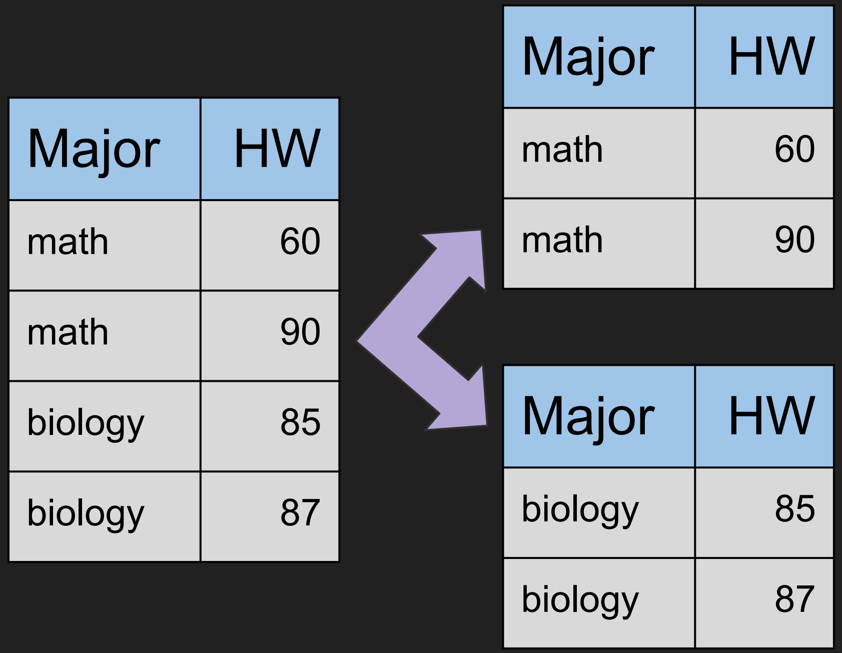
Major	hw_m
math	75

Major	hw_m
biology	86

Major	hw_m
math	75
biology	86

group_by() - split data by groups

```
group_by(., Major)
```



Major	HW
math	60
math	90
biology	85
biology	87

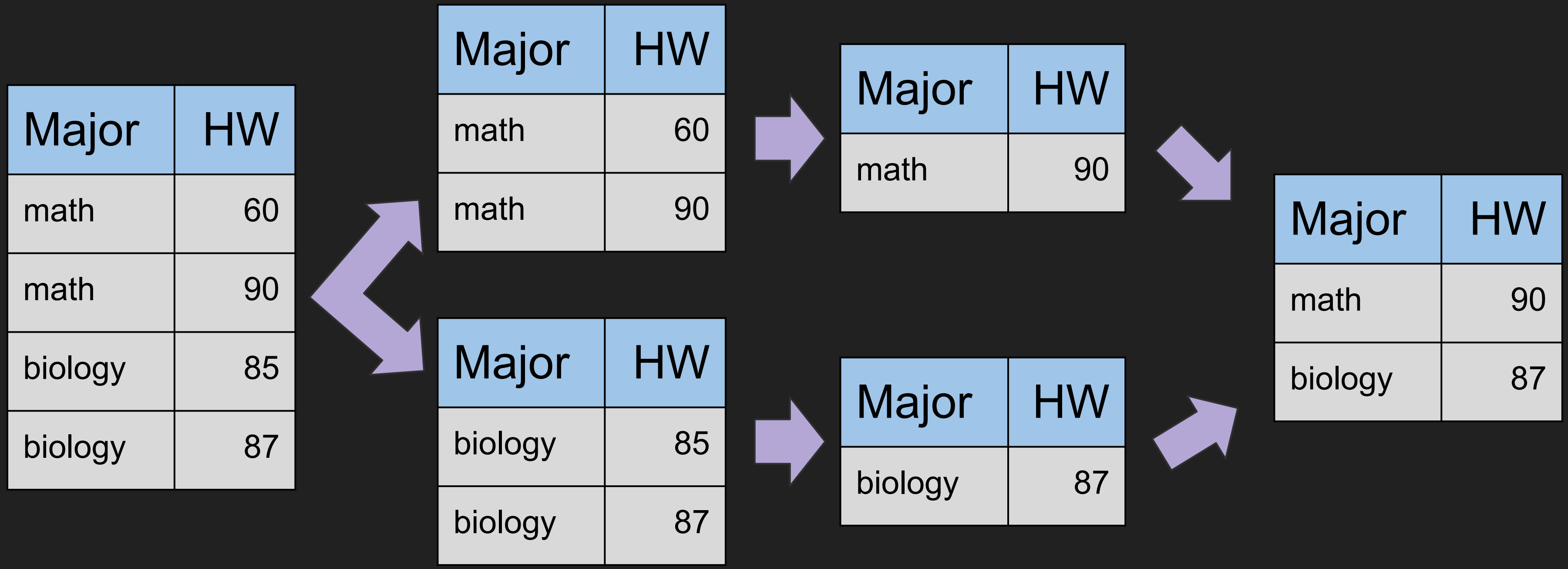
Major	HW
math	60
math	90

Major	HW
biology	85
biology	87

group_by() - split data by groups

Take top HW value

```
group_by(., Major) %>% top_n(1, HW)
```



Major	HW
math	60
math	90
biology	85
biology	87

Major	HW
math	60
math	90

Major	HW
biology	85
biology	87

Major	HW
math	90

Major	HW
biology	87

Major	HW
math	90
biology	87

summarize() and group_by() - Titanic Exercises

1. What was the average `Fare` cost?

```
> titanic %>% summarize(FareMean = mean(______))
```

2. What was the average `Age` for each `Pclass`? Store the results in `ClassAge`.

```
> ClassAge <- titanic %>%
```

```
> group_by(______) %>% summarize(_____ = ____(____))
```

3. What percent of each `Pclass` survived? (Hint: sum() and n() are helpful).

summarize() and group_by() - Titanic Exercise Solutions

1. What was the average `Fare` cost?

```
> titanic %>% summarize(FareMean = mean(Fare))
```

summarize() and group_by() - Titanic Exercise Solutions

1. What was the average `Fare` cost?

```
> titanic %>% summarize(FareMean = mean(Fare))
```

2. What was the average `Age` for each `Pclass`? Store the results in `ClassAge`.

```
> ClassAge <- titanic %>%
> group_by(Pclass) %>%
> summarize(AgeM = mean(Age))
```

summarize() and group_by() - Titanic Exercises

3. What percent of each `Pclass` survived? (Hint: sum() and n() are helpful).

```
> titanic %>%
>   group_by(Pclass) %>%
>   summarize(pct = sum(Survived)/n())
```

The Fundamental Functions

select	Select columns
filter	Filter rows
mutate	Create new variables
summarize	Compute summary variable
group_by	Split data by groups

The Fundamental Functions

select	Select columns
filter	Filter rows
mutate	Create new variables
summarize	Compute summary variable
group_by	Split data by groups
left_join	Combine datasets

left_join() - combine two datasets

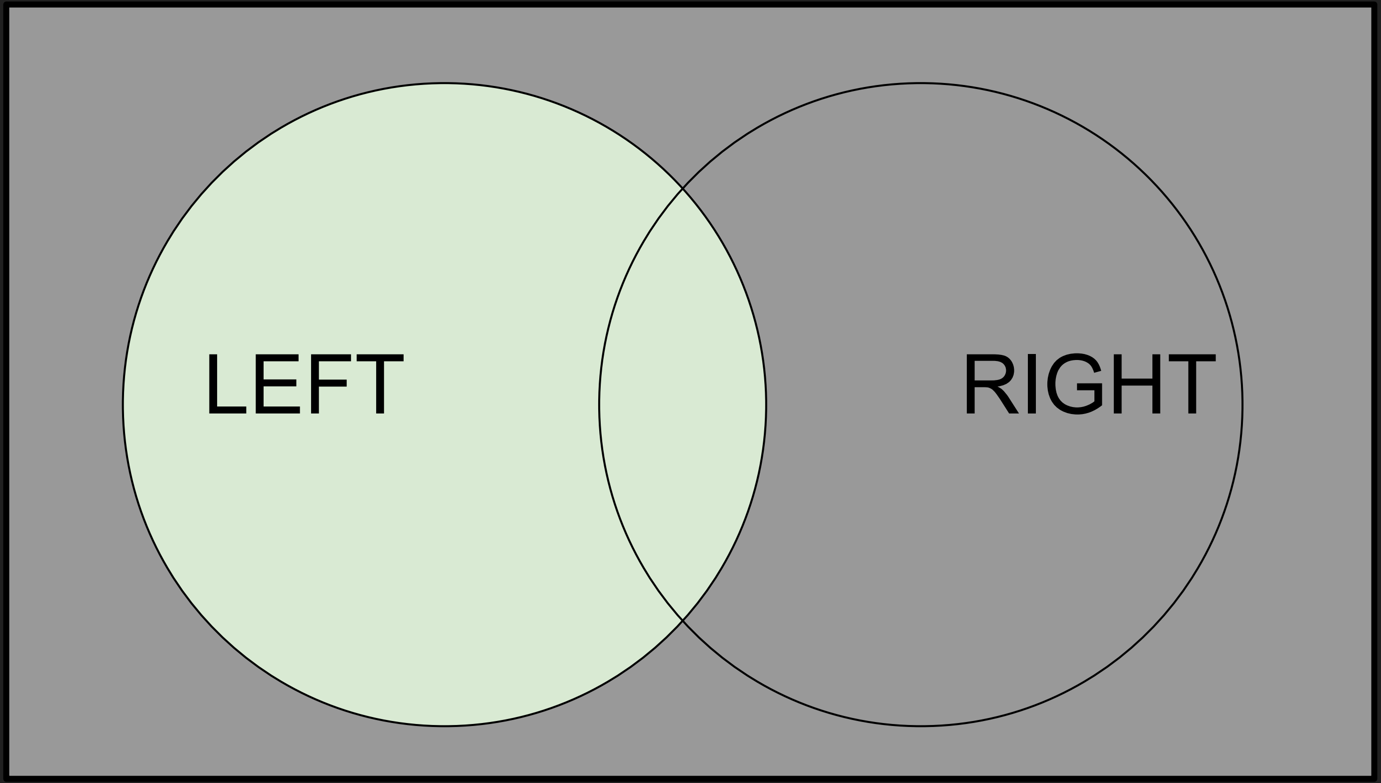
$$\texttt{left\_join(}\underbrace{\texttt{df\_left, df\_right,}}_{\boxed{\text{The two datasets to combine together}}}\texttt{ ...)}$$

Join two tables, matching on a key column

left_join() - combine two datasets

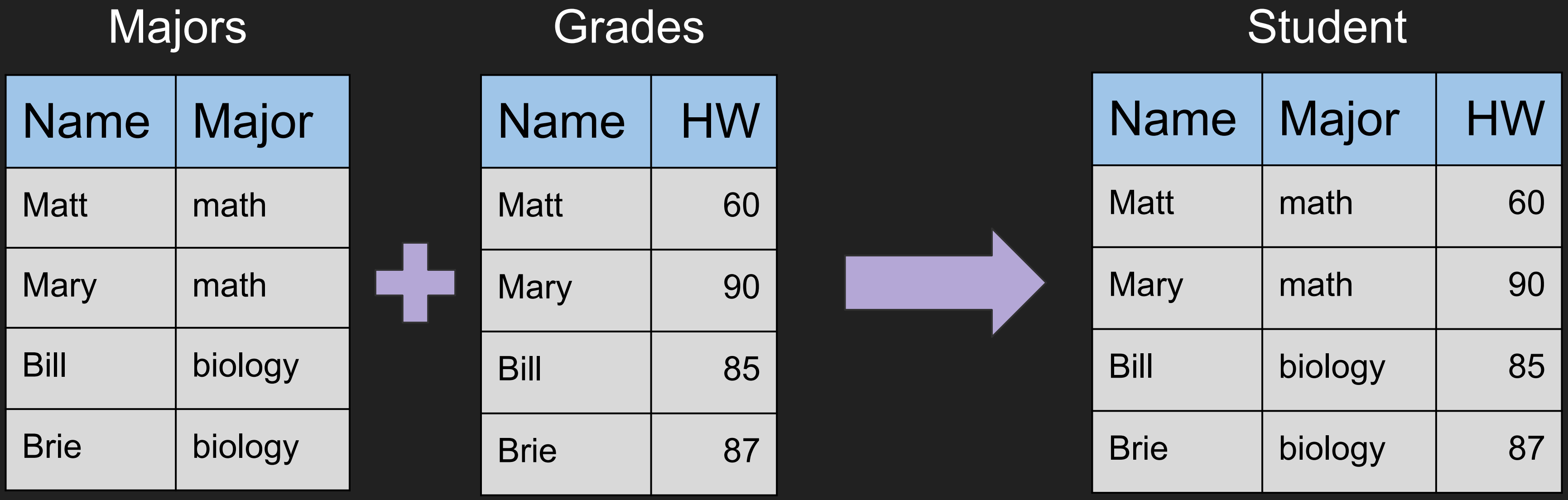
All values in the left dataset are kept

Only matching values in the right dataset are kept



left_join() - combine two datasets

Student <- left_join(Majors, Grades)



Majors

Name	Major
Matt	math
Mary	math
Bill	biology
Brie	biology

Grades

Name	HW
Matt	60
Mary	90
Bill	85
Brie	87

Student

Name	Major	HW
Matt	math	60
Mary	math	90
Bill	biology	85
Brie	biology	87

left_join() - combine two datasets

```
Student <- left_join(Majors, Grades)
```

Majors

Name	Major
Matt	math
Mary	math
Bill	biology
Brie	biology

Grades

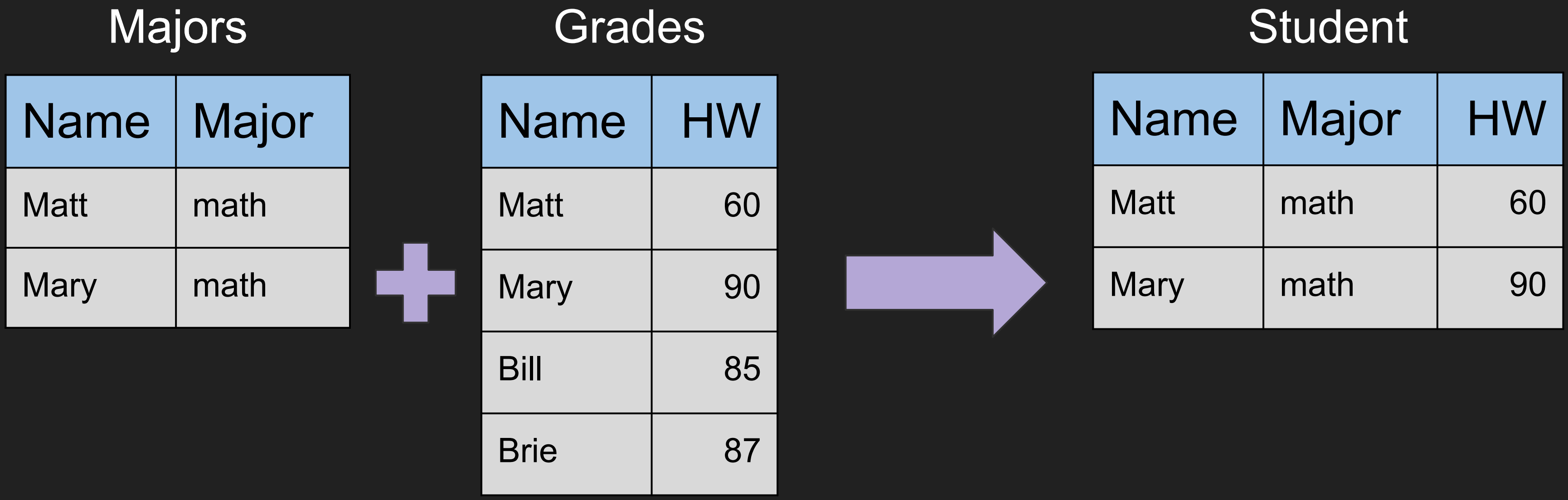
Name	HW
Matt	60
Mary	90

Student

Name	Major	HW
Matt	math	60
Mary	math	90
Bill	biology	*NA*
Brie	biology	*NA*

left_join() - combine two datasets

Student <- left_join(Majors, Grades)



Majors

Name	Major
Matt	math
Mary	math

Grades

Name	HW
Matt	60
Mary	90
Bill	85
Brie	87

Student

Name	Major	HW
Matt	math	60
Mary	math	90

left_join() - combine two datasets

If matching columns have different names

Student <- left_join(Grades, Majors, by = c("NX" = "NY"))

Majors

NX	Major
Matt	math
Mary	math
Bill	biology
Brie	biology

Grades

NY	HW
Matt	60
Mary	90
Bill	85
Brie	87

Student

NX	Major	HW
Matt	math	60
Mary	math	90
Bill	biology	85
Brie	biology	87

Other joins

Function	Description
left_join	Keep all left
right_join	Keep all right
inner_join	Match left and right
full_join	Keep all values
nest_join	Create list with match
anti_join	Keep non-matching

left_join() - Titanic Exercise

1. Split the titanic dataset accordingly:

```
> titanic1 <- titanic %>% select(1:3)
```

```
> titanic2 <- titanic %>% select(3:5)
```

2. Combine the datasets back together based on `Name`

```
> titanic3 <- left_join(_____, _____)
```

left_join() - Titanic Exercise

1. Split the titanic dataset accordingly:

```
> titanic1 <- titanic %>% select(1:3)
```

```
> titanic2 <- titanic %>% select(3:5)
```

2. Combine the datasets back together based on `Name`

```
> titanic3 <- left_join(titanic1, titanic2)
```

Exporting Data

Name in R

```
write_csv(dataset, file = “dataset.csv”)
```

Output name and path to location

NOTE: MUST INCLUDE .CSV

Exporting Data - Titanic Exercises

1. Store the first four columns in `TitanicReduced`

```
> TitanicReduced <- titanic %>% select(____)
```

2. Export the previously created dataset

```
> write_csv(_______, “RTitanic.csv”)
```

Exporting Data - Titanic Exercise Answers

1. Store the first four columns in `TitanicReduced`

```
> TitanicReduced <- titanic %>% select(1:4)
```

2. Export the previously created dataset

```
> write_csv(TitanicReduced, “RTitanic.csv”)
```

Other Tips

Working with dates/times? Use the **`lubridate` package!**

Skim through the data wrangling cheat sheet to see what is possible.

Do not reinvent the wheel. Use google before you try to code something fancy.

Have fun!