Data Analysis and Visualization with R for Social Scientists

Tugba Ozturk, PhD

Washington University in St Louis

Workshop @ GSDE 2020, Concordia University

R is a commonly-used programming language in many scientific disciplines for statistical analysis and for its powerful data science packages.

- ♥ simple syntax
- ♥ versatility
- ♥ well-developed packages
- ♥ inclusive & supportive community

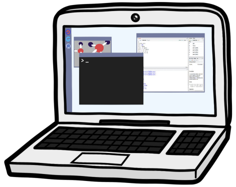
Agenda

- ✓ Navigate through a project folder using **RStudio**
- ✓ Become familiar with good coding practices and **R terminology**
- ✓ **Install** and use **R packages** that are commonly used in data science
- ✓ **Read** data files with R
- ✓ Inspect, **clean** and modify data sets
- ✓ Perform simple **statistical analysis**
- ✓ Generate publication-quality **graphs**

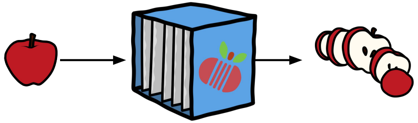
Q/A

Programming language

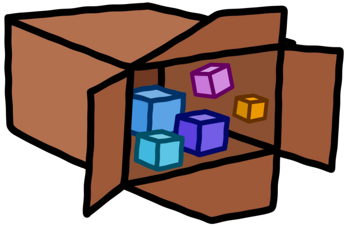
Integrated Development Environment (IDE)



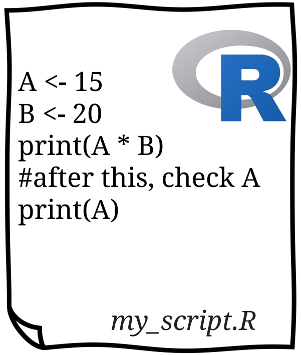
Function



R package



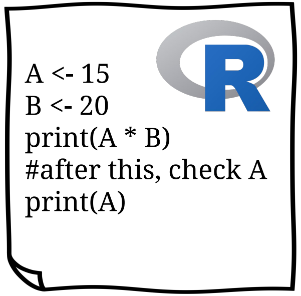
R script



```
A <- 15
B <- 20
print(A * B)
#after this, check A
print(A)
```

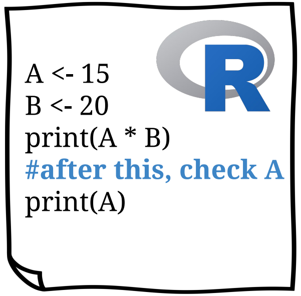
my_script.R

Execute



```
A <- 15
B <- 20
print(A * B)
#after this, check A
print(A)
```

Comment



```
A <- 15
B <- 20
print(A * B)
#after this, check A
print(A)
```

Errors!!!

Syntax errors

Invalid code that R doesn't understand

Semantic errors

Valid code that R understands, but it doesn't do what you intended

Logical errors

Valid code that R understands; it does what you intended; but the output is wrong…

Bare minimum

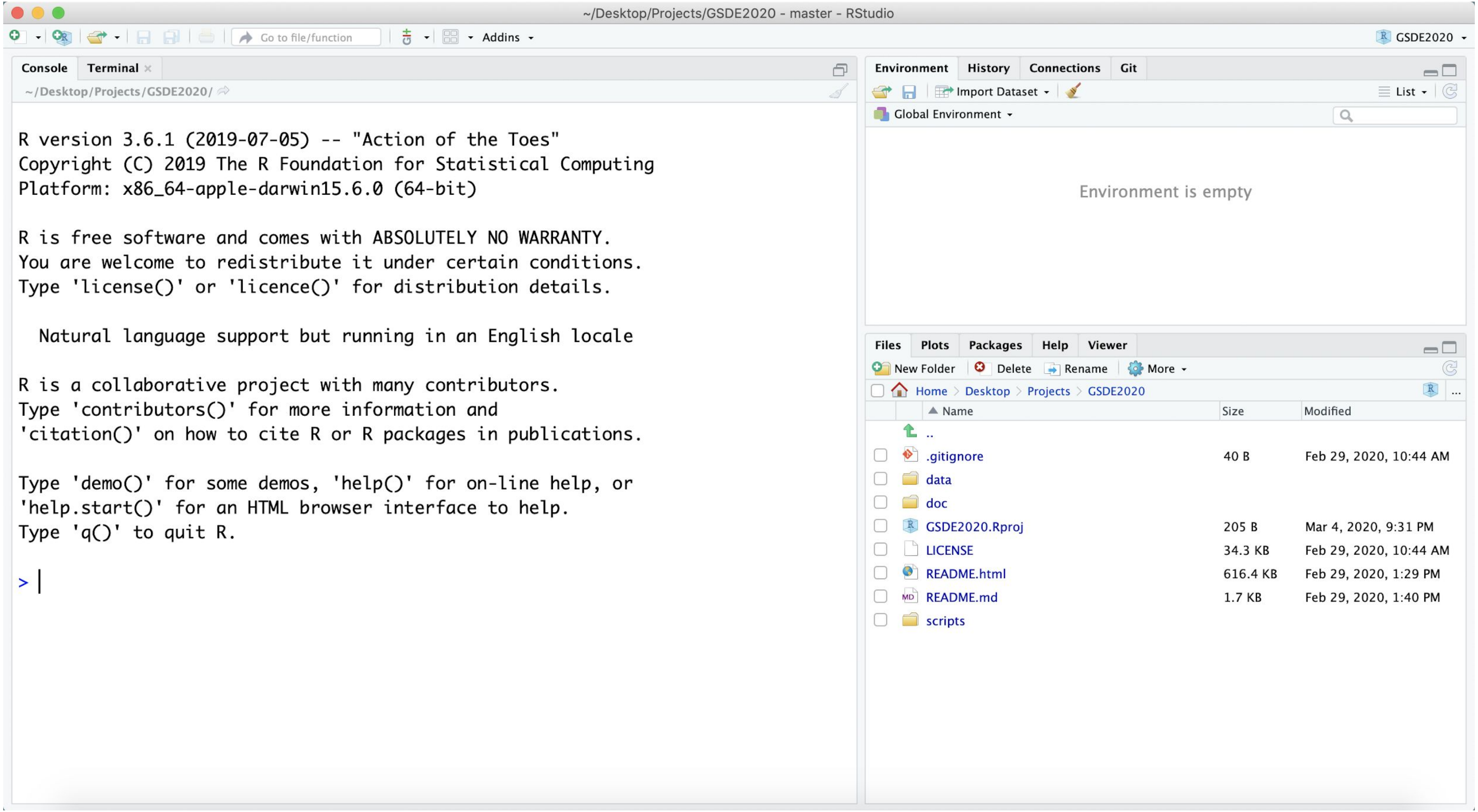
Single variable (numeric or character)

Vectors (list of variables)

Factors (for categorical values)

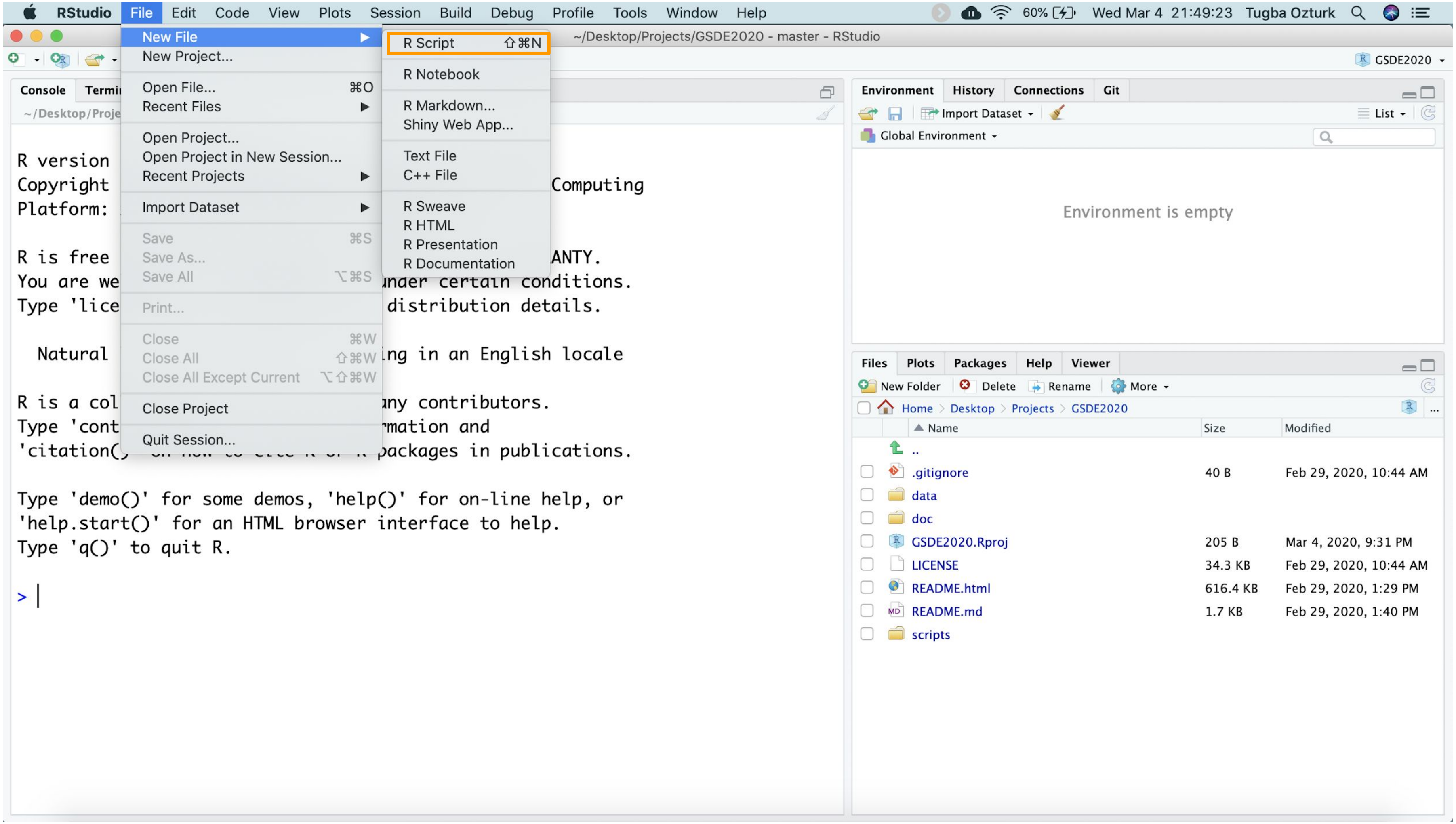
Data frames (for tabular data)

```
name_a_variable <- value(s)
```



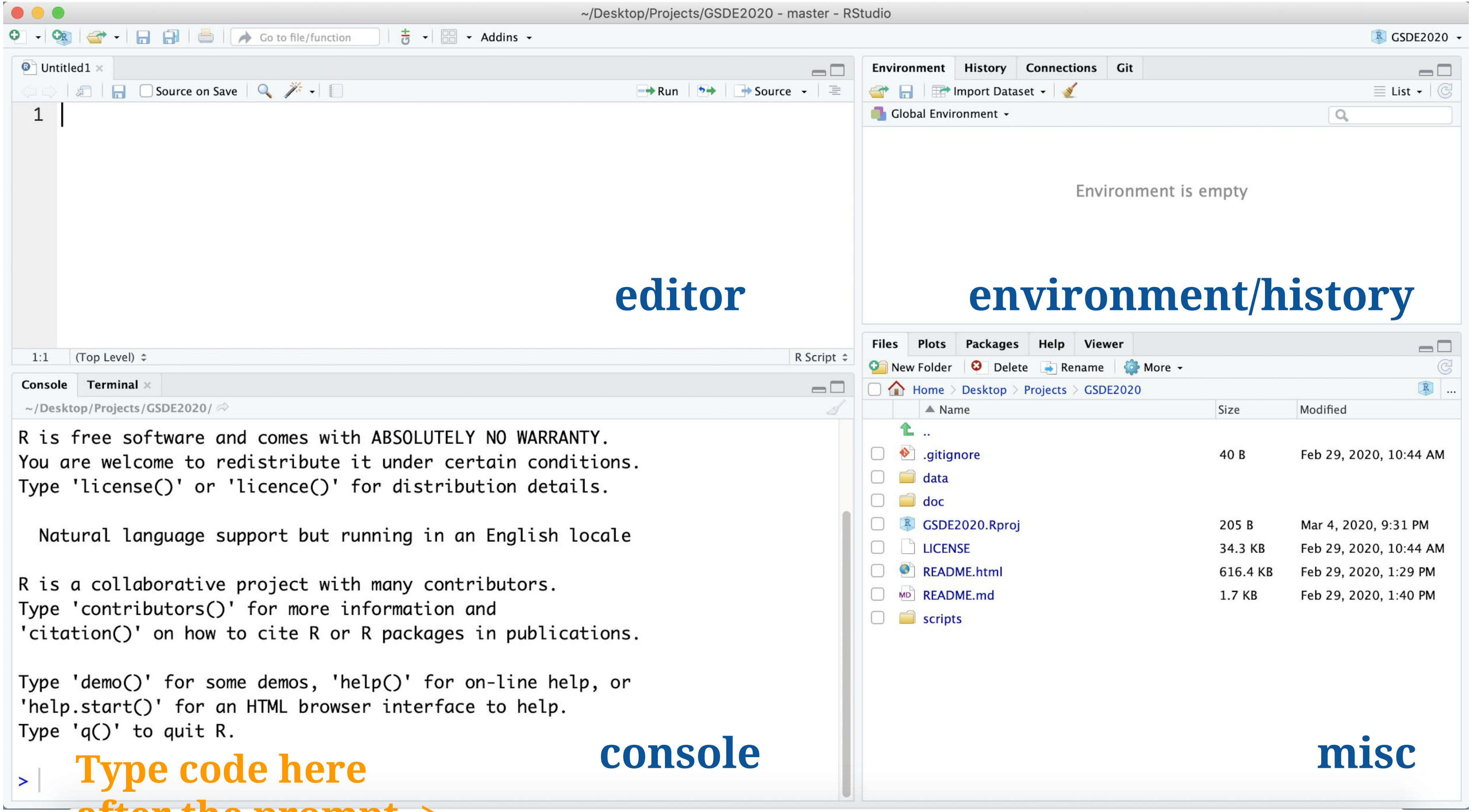
```
R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

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	40 B	Feb 29, 2020, 10:44 AM
data		
doc		
GSDE2020.Rproj	205 B	Mar 4, 2020, 9:31 PM
LICENSE	34.3 KB	Feb 29, 2020, 10:44 AM
README.html	616.4 KB	Feb 29, 2020, 1:29 PM
README.md	1.7 KB	Feb 29, 2020, 1:40 PM
scripts		



```
R version                                  
Copyright                                  Computing
Platform: 

R is free                                  ANTY.
You are we                                 under certain conditions.
Type 'lice                                 distribution details.

  Natural                                  ing in an English locale

R is a col                                 any contributors.
Type 'cont                                 rmation and
'citation()                                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	40 B	Feb 29, 2020, 10:44 AM
data		
doc		
GSDE2020.Rproj	205 B	Mar 4, 2020, 9:31 PM
LICENSE	34.3 KB	Feb 29, 2020, 10:44 AM
README.html	616.4 KB	Feb 29, 2020, 1:29 PM
README.md	1.7 KB	Feb 29, 2020, 1:40 PM
scripts		



```
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	40 B	Feb 29, 2020, 10:44 AM
data		
doc		
GSDE2020.Rproj	205 B	Mar 4, 2020, 9:31 PM
LICENSE	34.3 KB	Feb 29, 2020, 10:44 AM
README.html	616.4 KB	Feb 29, 2020, 1:29 PM
README.md	1.7 KB	Feb 29, 2020, 1:40 PM
scripts		

editor

environment/history

console

misc

Type code here
after the prompt, >

First steps

```
2
2+4
2**3
fav_colors <- c("blue","red")
print(fav_colors)
```

Create an object named **x** containing the value 1.5

```
x <- 1.5
```

Create an object named **x** containing the value 1.5

```
x <- 1.5
```

RStudio’s shortcut for the assignment operator: **Alt+-**
Try **Alt+Shift+K**

The name of an object is the reference to a value — the assignment arrow creates a binding from the name to the object.

Create an object named **x** containing the value “b”

```
x <- "b"
```

HOW TO NAME AN OBJECT?

- Do not start with a number or “_”
- Do not use “ ”
- Do not use a reserved word
  (Check by typing **`?Reserved`** after the prompt)
- Try to stick with simple english words relevant to your variable/data.

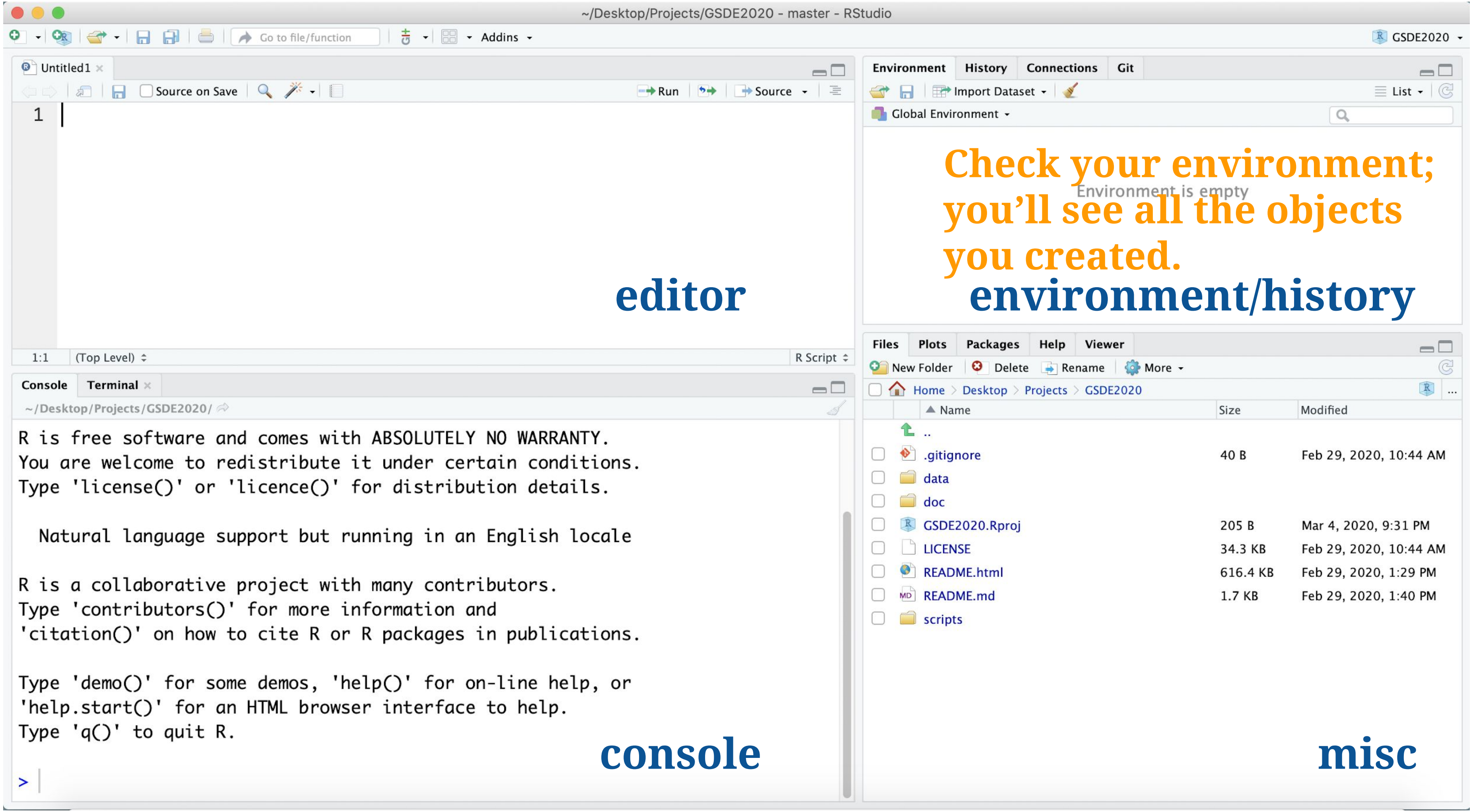
Calling functions

```
function_name(arg1=value1,arg2=value2,...)
x <- seq(3,12)
y <- seq(3,12,length.out=5)
```

Calling functions

```
function_name(arg1=value1,arg2=value2,...)
x <- seq(3,12)
y <- seq(3,12,length.out=5)
```

?function_name — for more details about a function

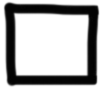


Check your environment; you'll see all the objects you created.

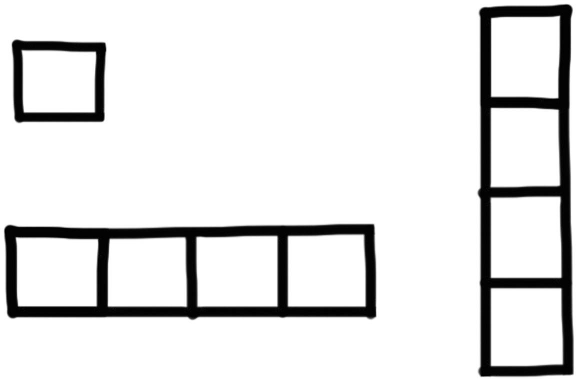
```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  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	40 B	Feb 29, 2020, 10:44 AM
data		
doc		
GSDE2020.Rproj	205 B	Mar 4, 2020, 9:31 PM
LICENSE	34.3 KB	Feb 29, 2020, 10:44 AM
README.html	616.4 KB	Feb 29, 2020, 1:29 PM
README.md	1.7 KB	Feb 29, 2020, 1:40 PM
scripts		

Vectors & Data Frames



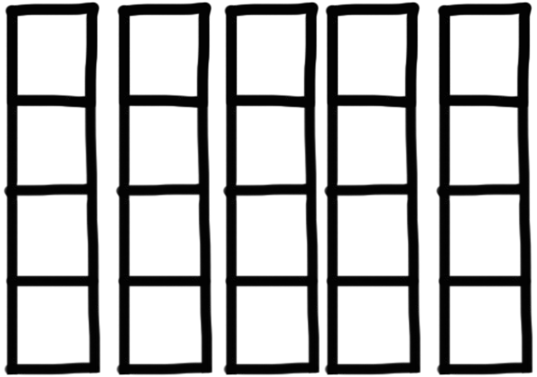
Vectors & Data Frames



```
vector_b - vector_a
vector_c[vector_c>5]
unique(vector_c)
vector_c[5]
vector_c[c(5,2,1)]
length(vector_c)
vector_a[-2]
```

```
vector_a <- c(4, 8, 2, 0)
vector_b <- seq(22, 2, -4)
vector_c <- c(vector_a, vector_b)
print(3*vector_b)
```

Vectors & Data Frames



```
vector_a <- c(34, 32, 32)
vector_b <- c(2, 0, 1)
my_data <- data.frame(age=vector_a,
                      nkids=vector_b)
```

```
dim(my_data)
colnames(my_data)
my_data[1,3]
my_data[,2]
my_data[c(3,1),]
```

Installing and loading R packages

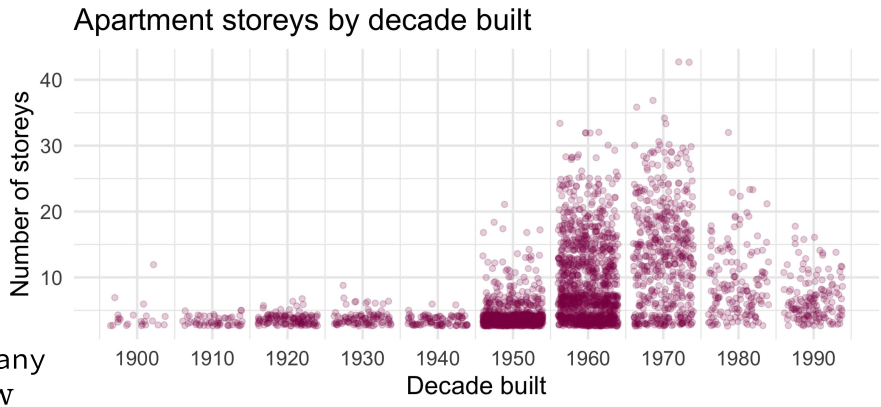
```
install.packages("tidyverse")
library("tidyverse")
```

Hands-on!

https://github.com/tnozturk/GSDE2020

Exercise

- ✓ Read the data file named *toronto_apartment_building_evaluation.csv*
- ✓ Inspect the data set using R
- ✓ Set the capitalization of the column names to lowercase (hint:`janitor::clean_names`)
- ✓ Change `N/A` values to `NA` so that R understands them as missing values (hint:`naniar::replace_with_na_all`)
- ✓ Inspect the data using basic graphs (For example, plot a histogram graph of the variable `year_built`)
- ✓ Create a data set for all data from 1900s (hint:`dplyr::filter`)
- ✓ Figure out the mean average of the stairwells and how many missing values exist in the `laundry_rooms` variable for the new data set.
- ✓ Create a new variable named `decade_built` using the `year_built` variable (hint: `dplyr::mutate`)
- ✓ Plot the graph given on the right only for the buildings with 3 or more storeys and save it as a PNG file (20 cm x 10 cm) (hint:`geom_jitter`, `ggsave`)


Apartment storeys by decade built

Source: Toronto Open Data

This exercise is adapted from Sharla Gelfand's talk: https://github.com/sharlagelfand

A list of resources

✓ https://www.rforexcelusers.com

✓ useR!2017-2019 (and soon useR! 2020 STL) videos

✓ Rstudio’s YouTube Channel

✓ https://education.rstudio.com/learn

✓ https://stat545.com

✓ https://www.rforexcelusers.com

✓ https://datacarpentry.org/r-socialsci

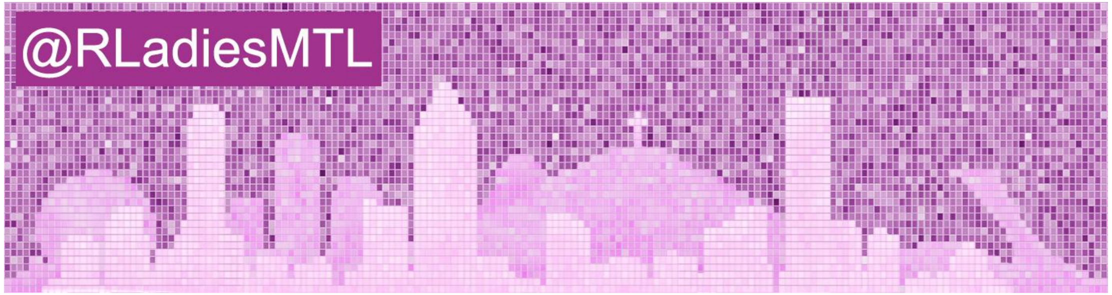
✓ https://software-carpentry.org/lessons

Tips for Excel/SPSS users

If you need to read SPSS, SAS and Stata files with R, check the package **haven**.

If you need to read Excel files with R, check the following R packages: **readxl**, **xlsx** and **xlsReadWrite** (Windows only).

Join us!



https://www.meetup.com/rladies-montreal