Oct 19, 2022

Introduction to Data Analysis in R (Part I)

Your first step in learning how to make your data work for you!

SECTION I: WHAT IS R?

Section I: What is R?

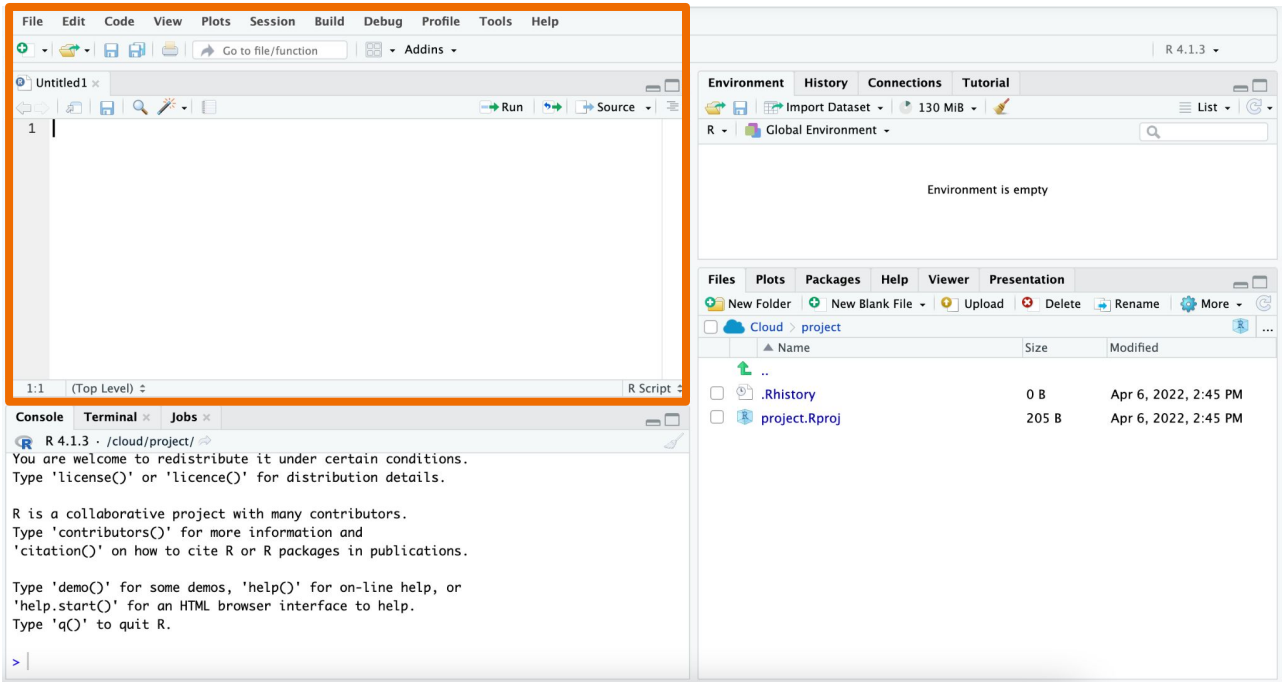
R is a programming language for statistical computing and graphics that is used extensively by researchers, statisticians and data scientists to clean, analyze, and graph their data.

It can be used to:

- Import data from your computer, websites and databases
- Clean the data by organizing it into matrices, data frames, or tables
- Analyze the data using statistical tests or graphs
- Communicate your results

The Different Components That Make Up R

A



File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | R 4.1.3

Untitled1

Run | Source

1

1:1 (Top Level) R Script

Console Terminal Jobs

R 4.1.3 · /cloud/project/

```
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

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

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

>
```

B

Environment History Connections Tutorial

Import Dataset | 130 MiB | List

R | Global Environment

Environment is empty

Files Plots Packages Help Viewer Presentation

New Folder | New Blank File | Upload | Delete | Rename | More

Cloud > project

Name	Size	Modified
..		
.Rhistory	0 B	Apr 6, 2022, 2:45 PM
project.Rproj	205 B	Apr 6, 2022, 2:45 PM

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | R 4.1.3

Untitled1

Run | Source

1

C

1:1 (Top Level) R Script

Console Terminal Jobs

R 4.1.3 · /cloud/project/

```
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

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

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

>
```

Environment History Connections Tutorial

Import Dataset | 130 MiB | List

R | Global Environment

Environment is empty

Files Plots Packages Help Viewer Presentation

New Folder | New Blank File | Upload | Delete | Rename | More

Cloud > project

Name	Size	Modified
..		
.Rhistory	0 B	Apr 6, 2022, 2:45 PM
project.Rproj	205 B	Apr 6, 2022, 2:45 PM

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | R 4.1.3

Untitled1

Run | Source

1

1:1 (Top Level) R Script

Console Terminal Jobs

R 4.1.3 · /cloud/project/

```
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

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

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

>
```

D

Environment History Connections Tutorial

Import Dataset | 130 MiB | List

R | Global Environment

Environment is empty

Files Plots Packages Help Viewer Presentation

New Folder | New Blank File | Upload | Delete | Rename | More

Cloud > project

Name	Size	Modified
..		
.Rhistory	0 B	Apr 6, 2022, 2:45 PM
project.Rproj	205 B	Apr 6, 2022, 2:45 PM

SECTION II: MAKING THE BEST OF R MARKDOWN & VARIABLES

Section II: Variable Assignments

- A variable is a named storage space that we can manipulate and mutate using code in R.
- We often name variables something meaningful to our program for readability

- E.g. Math: x = 5

  R: x<- 5

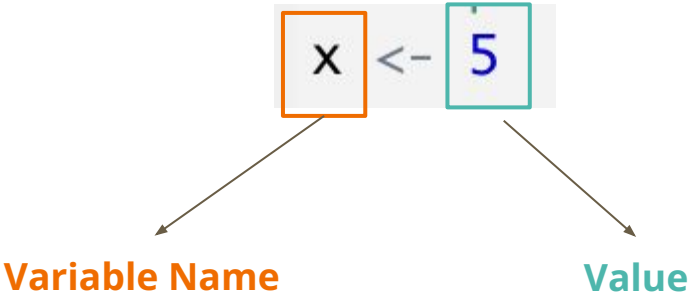
Section II: Variable Assignments

- To save values and result to access later
- To simplify your code so one single line does not run too long
- Variable is defined using **"<-"**

```
x <- 5
```

Section II: Variable Assignments

- To save values and result to access later
- To simplify your code so one single line does not run too long
- Variable is defined using **"<-"**
-



Section II: Variable Assignments

- Three Types of Variables
  - Integer/Numeric → Numbers
  - Character → Words/Text
  - Logical → TRUE/FALSE

SECTION III: DATA OBJECTS IN R

Section III: Data Objects – Vectors

- Vectors: a sequence of value of the same data type

1
2
3
4

Section III: Data Objects – Vectors

- Vectors: a sequence of value of the same data type

1
2
3
4

“Sun”
“Mon”
“Tues”
“Wed”

Section III: Data Objects – Vectors

- Vectors: a sequence of value of the same data type

1	“Sun”	TRUE
2	“Mon”	TRUE
3	“Tues”	FALSE
4	“Wed”	TRUE

Section III: Data Objects – Vectors

- Vector Calculation

1
2
3
4

multiply $\boxed{2}$ = ?

Section III: Data Objects – Vectors

- Vector Calculation

1
2
3
4

multiply

2

=

1*2
2*2
3*2
4*2

=

2
4
6
8

Section III: Data Objects – Data Frames

- Vectors: 1D vertical “list”
- Data Frame: 2D : row component + column component

2
3
4

“Mon”
“Tues”
“Wed”

Section III: Data Objects – Data Frames

- Vectors: 1D vertical “list”
- Data Frame: 2D : row component + column component

2	“Mon”
3	“Tues”
4	“Wed”

Section III: Data Objects – Data Frames

- Vectors: 1D vertical “list”
- Data Frame: 2D : row component + column component

ID	Day
2	“Mon”
3	“Tues”
4	“Wed”

Column Names

SECTION IV: DATA SUBSETTING

Section IV: Subsetting Vectors

- Subsetting allows you to access specific elements in your data objects
- We will use variable_name**[index]** to access the elements

“Sun”
“Mon”
“Tues”
“Wed”

Section IV: Subsetting Vectors

- Subsetting allows you to access specific elements in your data objects
- We will use variable_name**[index]** to access the elements

Index	
1	“Sun”
2	“Mon”
3	“Tues”
4	“Wed”

Section IV: Subsetting Data Frames(with index)

- Subset by Row
- Subset by Column
- Subset by Row & Column

ERC_fellow[,]

Row

Column

ID	UNI	name
1	qa2116	Grace
2	cy3905	Winnie

Section IV: Subsetting Data Frames(with index)

- **Subset by Row**
- Subset by Column
- Subset by Row & Column

ERC_fellow[■ ,]

Row Index

1
2

ID	UNI	name
1	qa2116	Grace
2	cy3905	Winnie

Section IV: Subsetting Data Frames(with index)

- Subset by Row
- **Subset by Column**
- Subset by Row & Column

ERC_fellow[,]

Column Index

1	2	3

ID	UNI	name
1	qa2116	Grace
2	cy3905	Winnie

Section IV: Subsetting Data Frames(with index)

- Subset by Row
- Subset by Column
- **Subset by Row & Column**

ERC_fellow[,]

Column Index

1	2	3

Row Index

1
2

ID	UNI	name
1	qa2116	Grace
2	cy3905	Winnie

Section IV: Subsetting Data Frames(with logical operator)

- Logical Operator definition: test conditions and output either TRUE or FALSE
- Logical Operators:
  - ==
  - >/<
  - !=

Section IV: Subsetting Data Frames(with logical operator)

- Logical Operator definition: test conditions and output either TRUE or FALSE

ERC_fellow[condition ,]

ID	UNI	name	
1	qa2116	Grace	TRUE
2	cy3905	Winnie	FALSE

SECTION V: CREATING YOUR OWN DATA ANALYTICS PROJECT in R!

Section V: Project Series Introduction

- Main Research Question: How much is the college wage premium?
- Leading Questions:
  - What is distribution of early_career_pay?
  - We can find out whether attending a **private vs public** colleges make in difference in **early_career_pay (mid_career_pay)**.
  - We can fit a linear regression analysis on **early_career_pay** versus **stem_percent** where early_career_pay is our response(dependent variable) and stem_percent is our independent variable
  - Finally, we want to figure out whether college tuition pays off. Is the difference between **mid_career_pay** and **tuition cost** positive?

Section V: Project Series Introduction

- Approach:
  - Summary Statistics
  - Clean the data if needed
  - Data Visualization to attempt at problem
  - Linear Regression to tackle the problem

Linear Regression

- Linear regression is a technique used to explain and understand the relationship between two quantitative variables.
- Fits a line of best fit to the scatter plot
- Y variable is the response variable
- X variable is the explanatory variable

$$Y = mX + b$$

m=coefficient of explanatory variable

B = y intercept

Linear Regression

A **hypothesis tes**t determines whether the relationship between the two variables is **statistically significant or not**.

Null hypothesis = no relationship between the two variables

Alternative hypothesis = statistical evidence must be strong enough (beyond reasonable doubt) that there is a relationship between the two variables

Hypothesis test on the regression produces:

P value = A number which quantifies how likely that the data occurred by random chance.

If $p < 0.05$, reject the null hypothesis in favor of the alternative

Linear Regression

Wrap Up

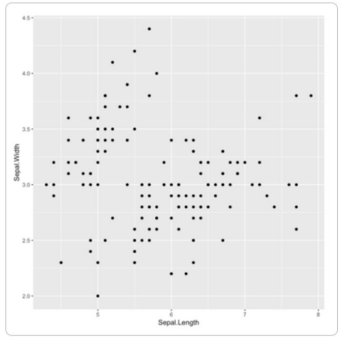
- R and R Markdowns
- Variable Assignment
- Data Objects
  - Vectors
  - Data Frame
  - Reading in csv files
- Subsetting
- Exploratory Data Analysis (EDA)

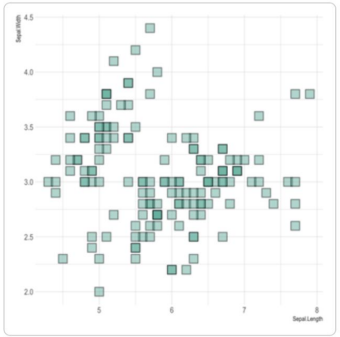
Want to do more with R?

YES!

Intermediate Workshop next Monday(04/18) at 4pm!

- Answer more questions about college wage premium
- Library and packages
- dplyr
- ggplots





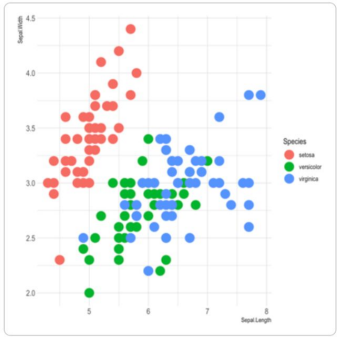


Image source:
https://r-graph-gallery.com/index.html

Thank you for attending this workshop!

Hope to see you again next week!

Workshop Presented by: Aarushi Sharma and Grace An(Qi An)

More information about ERC:
Website: https://erc.barnard.edu/
Walk-in Hours: https://erc.barnard.edu/visit-us
Email us at erc@barnard.edu