

Lecture 2a.

R Basics (1) (2)

POLI3148. Data Science in Politics and Public Administration

Dr. Haohan Chen

HKU-PPA

Outline

- Why R
- Resources that can help you learn
- R & RStudio intro
- R data science project setup
- R basics: Vector
- Start working with data

Why R?

- Powerful and user-friendly

- Data management: ★★★★★
- Exploratory data analysis: ★★★★★★
- Data visualization: ★★★★★
- Modeling
 - Statistical modeling: ★★★★★★
 - Machine learning: ★★★

- Vivid community

- Academia
- Industry




Where to get help

- CampusWire course forum
- Search engines
- StackOverflow
- Large Language Model
 - ChatGPT: <https://chatgpt.hku.hk/>
 - Ask questions about R and data science
 - Share your experience with the class (screenshots of chat and outcomes)

Troubleshooting is an important skill you will acquire here.

ChatGPT demo


Chat started

Haohan Chen

I am programming with R. Can you help me?

14/09/2023, 09:48:03

 **ChatGPT** (Tokens used 47)

Of course! What do you need help with in particular?

14/09/2023, 09:48:04

Haohan Chen

Can you write a program that import a .csv file to R?

14/09/2023, 09:48:24



Certainly! Here's an example program that imports a CSV file named "data.csv" located in the working directory of R:

```
r
# Set the file path to your CSV file
file_path <- "data.csv"

# Import the CSV file using the read.csv() function
data <- read.csv(file_path)

# Print the imported data to the console
print(data)
```

You can customize the file path to match the location of your CSV file. The `read.csv()` function is used to read in the CSV file and store it in the `data` variable. Finally, the `print()` function is used to display the imported data in the console.

14/09/2023, 09:48:24

Be specific about your request. For example, what packages you are using.

Haohan Chen

Can you do this using the tidyverse package?

14/09/2023, 09:48:38





Can you do this using the tidyverse package?

14/09/2023, 09:48:38



ChatGPT (Tokens used 503)

Yes, here's an example program using the tidyverse package to import a CSV file:

```
<pre style="background-color: rgb(43, 43, 43);margin-right: 15px;"><div class="pre-code-area"><code class="language-javascript" style="white-space: pre-wrap;">r
# Load the tidyverse package
library(tidyverse)

# Set the file path to your CSV file
file_path <- "data.csv"

# Import the CSV file using the read_csv() function
data <- read_csv(file_path)

# Print the imported data to the console
print(data)
</code></div></pre>
```

In this program, the `tidyverse` package is loaded using the `library()` function. The `read_csv()` function from the package is used to read in the CSV file and store it in the `data` variable. Finally, the `print()` function is used to display the imported data in the console. Note that `read_csv()` is similar to `read.csv()`, but provides additional functionality and data cleaning options.

It does not always get things right.

If it messes up, just ask it to try again.

14/09/2023, 09:48:38



The above is not correctly displayed. Can you try again?

14/09/2023, 09:49:05



ChatGPT (Tokens used 766)

Apologies for the formatting issue. Here's the correct example program using the tidyverse package to import a CSV file:

```
R
# Load the tidyverse package
library(tidyverse)

# Set the file path to your CSV file
file_path <- "data.csv"

# Import the CSV file using the read_csv() function
data <- read_csv(file_path)

# Print the imported data to the console
print(data)
```

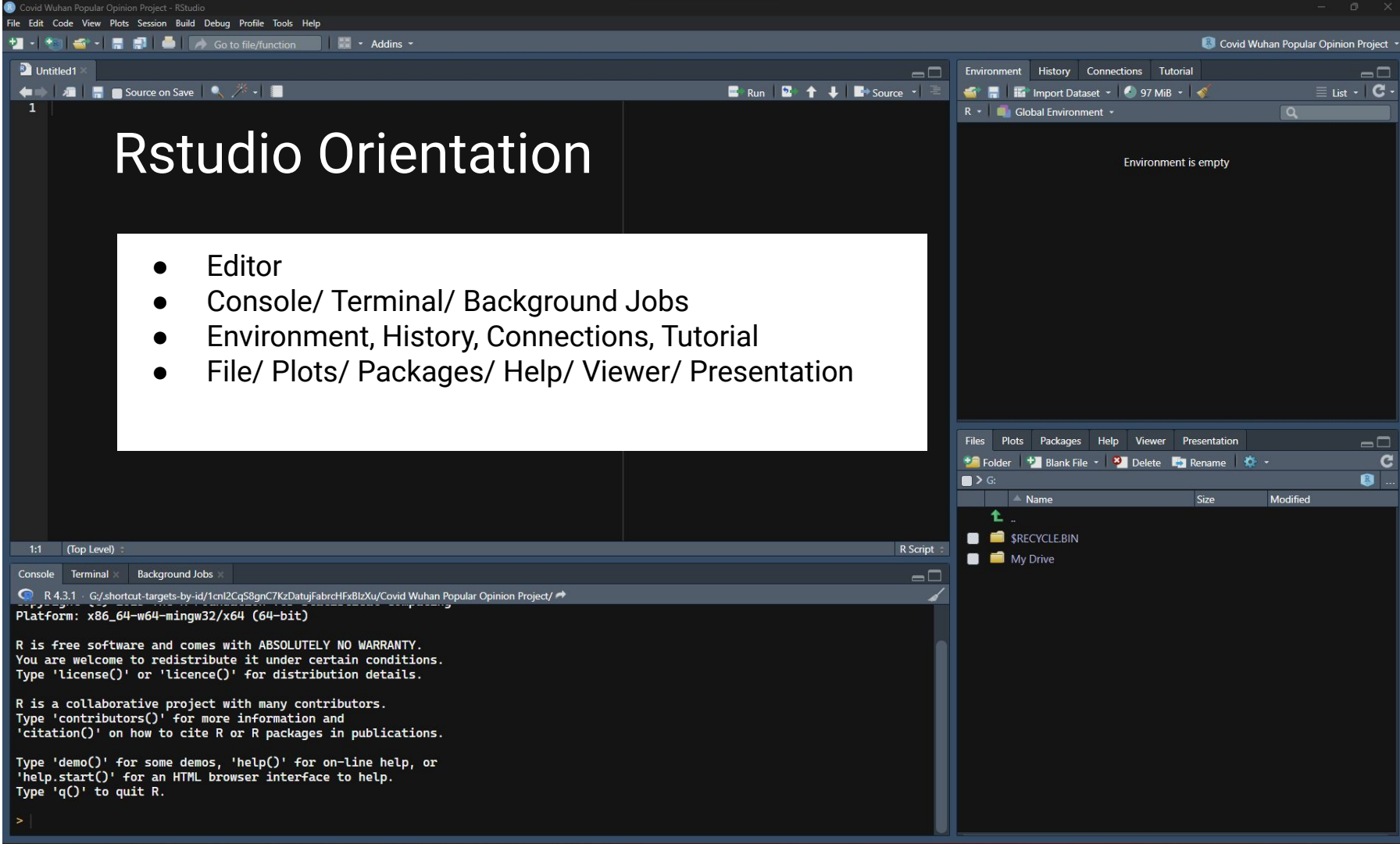
In this program, the `tidyverse` package is loaded using the `library()` function. The `read_csv()` function from the package is used to read in the CSV file and store it in the `data` variable. Finally, the `print()` function is used to display the imported data in the console. Note that `read_csv()` is similar to `read.csv()`, but provides additional functionality and data cleaning options.

14/09/2023, 09:49:05



R & RStudio intro

Douglas et al. Chapter 2



R Packages: R is a marketplace



R Packages

- Packages: Tools people to perform a variety of data science task
- R has a vivid community of developers offering tools.
- Package sources
 - **CRAN:** Packages on it need to pass some checks
Installation: `install.packages(...)`
 - **Github:** No check required. Anyone can put on a package
Installation: `remotes::install_github`
- Check if a package is installed
- Update package

The Tidyverse Package



R packages for data science

The tidyverse is an opinionated **collection of R packages** designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

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

R Data Science Project Setup

Douglas et al. Chapter 2

R Project: Why?

What Project does: Put **data**, **code**, and **outputs** in one place

- Stay organized
- Sharable and replicable analysis
 - Share one folder and that has everything
 - Audience can easily replicate the analysis

ALWAYS set up a *Project*

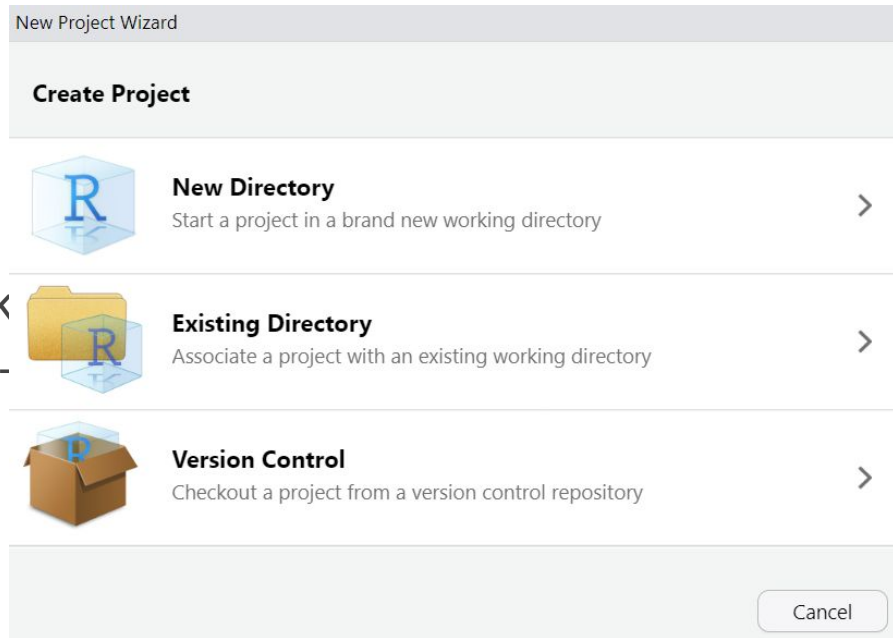
Set up a Project for EVERYTHING you do.

- Good practice
- Basic requirement for the *Assignment* and the *Replication Dossier*

R Project: How-To

Two Routes

- **New Directory:** Create R project folder under an existing folder
- **Existing Directory (recommended):** Make an R project under an existing directory - note: It will NOT remove anything under that existing folder
- **Version Control (next lecture)**



Create and examine Project

Use buttons in Rstudio to make folders

Functions

- `list.files()`
- `dir.create()`

Small point: Absolute path vs relative path

- Absolute path: Location from the system ROOT directory ("/")
- Relative path: Locations relative to the working directory

".." refers to the parent directory

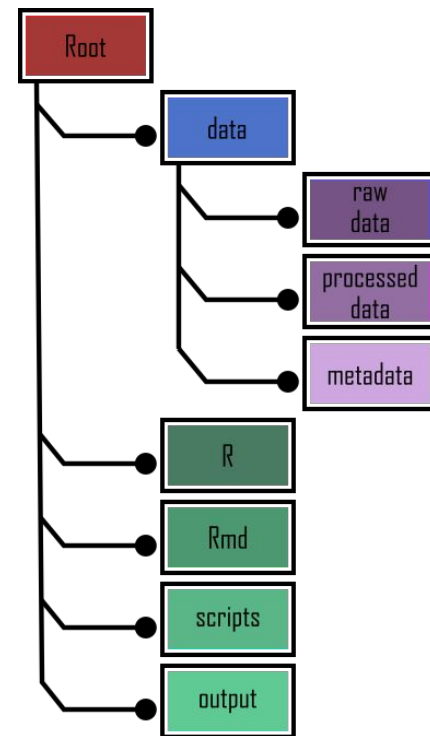
Project Structure

DO

- Store data, code, and outputs (figures, tables, documents) in separate folders
- Store raw and processed data in separate folders
- Use short but informative folder and file names
- Structure the project folder in ways you see fit

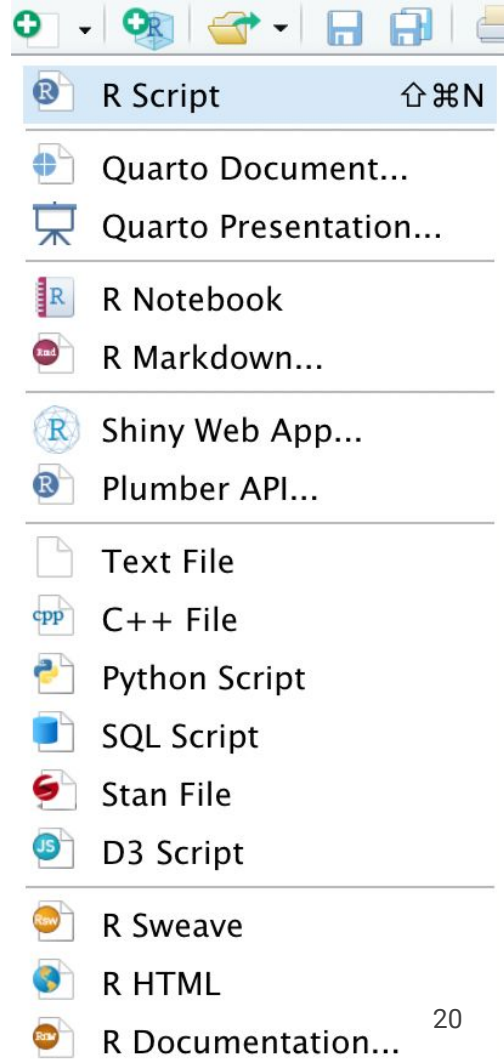
DON'T

- Don't put everything under the root folder



Create Files

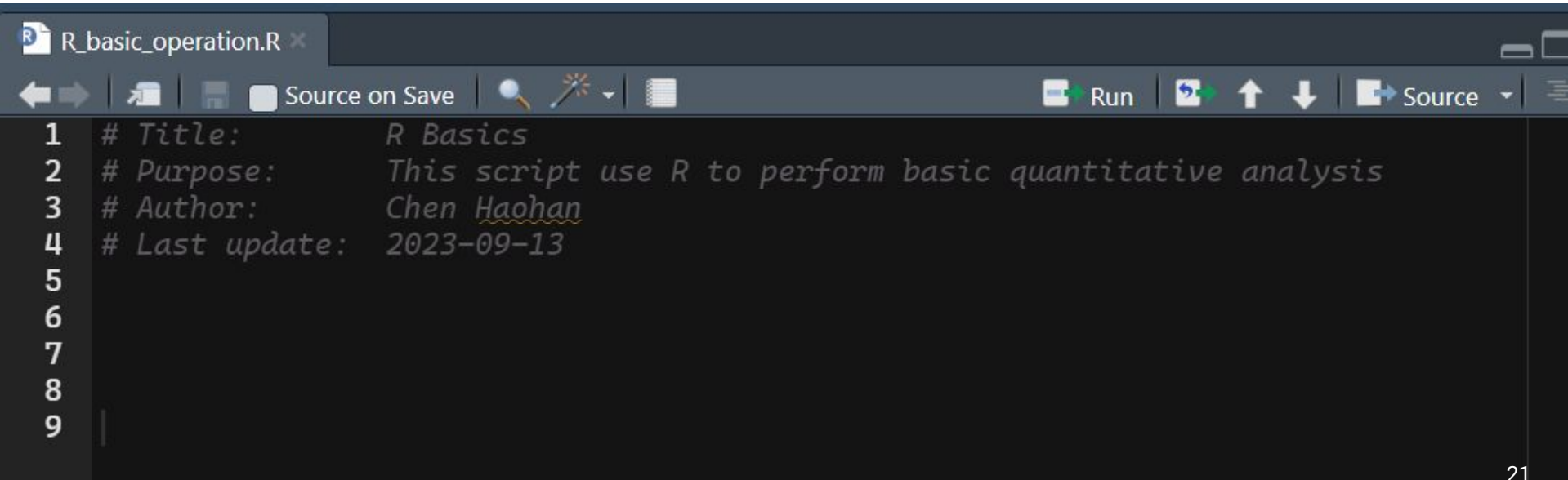
- Informative and concise file names
- **No space** in folder and file names. Use **underscores** in place of spaces. E.g., “raw_data”



Write your first R script

Header is essential for every R script (see 1.10 project documentation)

It summarises what a script does, who writes it, and when

A screenshot of an R script editor window. The title bar shows 'R_basic_operation.R'. The editor has a toolbar with icons for navigation, search, and execution. The script content is as follows:

```
1 # Title:      R Basics
2 # Purpose:    This script use R to perform basic quantitative analysis
3 # Author:     Chen Haohan
4 # Last update: 2023-09-13
5
6
7
8
9
```

R Data Structure 1: Vector

Douglas et al. Chapter 2

“Data Structure”


A data structure is a storage that is used to store and organize data. It is a way of arranging data on a computer so that it can be accessed and updated efficiently.

The Simplest data structure in R: Vector

How to create and manipulate the basic object, **vector**, in R

What is a vector?

“A vector is substantially **a list of variables**, and **the simplest data structure in R**. A vector consists of a collection of numbers, arithmetic expressions, logical values or character strings for example.”



R Data Structure 2: Data Types Matrix, Array, List

Douglas et al. Chapter 3


R Data types

- **Numeric:** Numbers that contain a decimal
- **Integer:** Whole number
- **Logical:** TRUE or FALSE
- **Character:** String values

Code with me

Check object's data type. `class()`; `typeof()`

Type	Logical test	Coercing
Character	<code>is.character</code>	<code>as.character</code>
Numeric	<code>is.numeric</code>	<code>as.numeric</code>
Logical	<code>is.logical</code>	<code>as.logical</code>
Factor	<code>is.factor</code>	<code>as.factor</code>
Complex	<code>is.complex</code>	<code>as.complex</code>



Vector is the simplest data structure.

Let's now move on to other more complicated data formats.

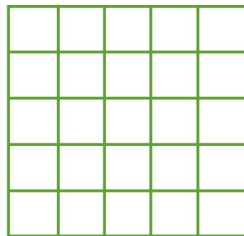
Scalar, vector, matrix, and array



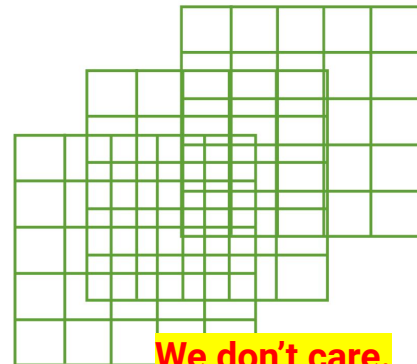
scalar



vector



matrix




array

List: A basket that everything can go in



R Data Structure 2: Data Frame

Douglas et al. Chapter 3



THE most
important data
structure for us.

What is a data frame?

A Data Frame is a data structure that organizes data into a **2-dimensional table** of **rows** and **columns**, much like a spreadsheet. DataFrames are one of the most common data structures used in modern data analytics because they are a flexible and intuitive way of storing and working with data.

Source: *DataBrick*

Data Frame in R

The diagram illustrates a Data Frame in R. It features a table with 7 rows and 6 columns. The columns are labeled *Name*, *Team*, *Number*, *Position*, and *Age*. The rows are indexed from 0 to 6. Annotations include: 'Columns' with arrows pointing to the column headers; 'Rows' with arrows pointing to the row indices; and 'Data' with a bracket pointing to the data cells. Some cells are highlighted with pink boxes: 'Jonas Jerebko', '8.0', 'Boston Celtics' (in row 3), 'PG' (in row 4), and 'NaN' (in row 5).

	<i>Name</i>	<i>Team</i>	<i>Number</i>	<i>Position</i>	<i>Age</i>
0	Avery Bradley	Boston Celtics	0.0	PG	25.0
1	John Holland	Boston Celtics	30.0	SG	27.0
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN
6	Evan Turner	Boston Celtics	11.0	SG	27.0

Source: <https://www.geeksforgeeks.org/r-data-frames/>

What is special about DataFrame as a data structure?

- **Dataframe vs Matrix:** Dataframe is different from matrix in that its columns can have different data types (a matrix's columns should have the same data type)
- **Dataframe vs List:** Dataframe is a special type of list. It is special in that its elements are **vectors** of the **same length**

A First Glance at Data Frame in R

- Create a data frame
- View the data frame
- Get basic information of the data frame
- Change basic information of the data frame
- Extract information from data frame by positions

In-class exercise: Meet the V-Dem Data

- Meet our example data: V-Dem data
- Download a folder of V-Dem data from [Moodle](#) to your **project folder**
- Unzip the folder
- Load the data saved in “V-Dem-CY-Full+Others-v13.csv”
(How? See Slide 8)
- Summarize basic information of the data
 - How many rows?
 - How many columns?
 - Pick a column of interest (consult codebook_v13.pdf)
 - What is this column about?
 - What is the data type?
 - Provide basic summary statistics
 - Are there missing values? How many of them are missing?
- Put your answers in your R scripts as comments (leading the lines with “#”)
- Submit your R script through Moodle (due by the end of Saturday)

Bonus Exercise (+1%): Tell Me Something about Yourself

Upload a **1-page** PDF document. Include:

- A photo of yourself
- A brief self-introduction

In addition:

- Let me know your prior experience with quantitative methods and programming (no worry if your answer is “none”)
 - Ask 1 question on “*data science in PPA*” that intrigues you the most
- 