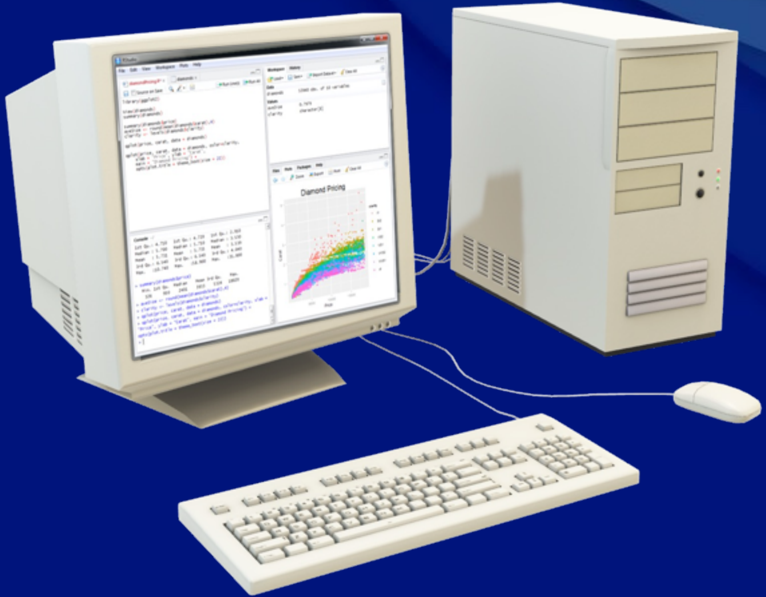


An Introduction to R

Workshop at John Jay College of Criminal Justice

By: Dilan Caro

JOHN JAY COLLEGE OF CRIMINAL JUSTICE
OFFICE FOR STUDENT RESEARCH AND CREATIVITY

About me

- BS Applied Mathematics: Cryptography and Data Science, Minor in Computer Science JJAY ‘23
- Pursuing MA in Statistics At Columbia University
- Conducting research in the field of Differential Privacy.

Preliminaries

- This 3 series workshop builds on my personal knowledge and other workshops or resources, below the sources for inspiration , graphics, examples:
    - https://cran.r-project.org/doc/manuals/r-release/R-intro.html
    - https://katrienantonio.github.io/intro-R-book/
    - https://unl-statistics.github.io/R-workshops/
    - https://intro2r.com/
    - https://bookdown.org/rdpeng/rprogdatascience/
    - https://r4ds.had.co.nz/
- Workshop materials on:
    - https://dilancaro.github.io/R-workshop-John-Jay/

Overview

Workshop 1:

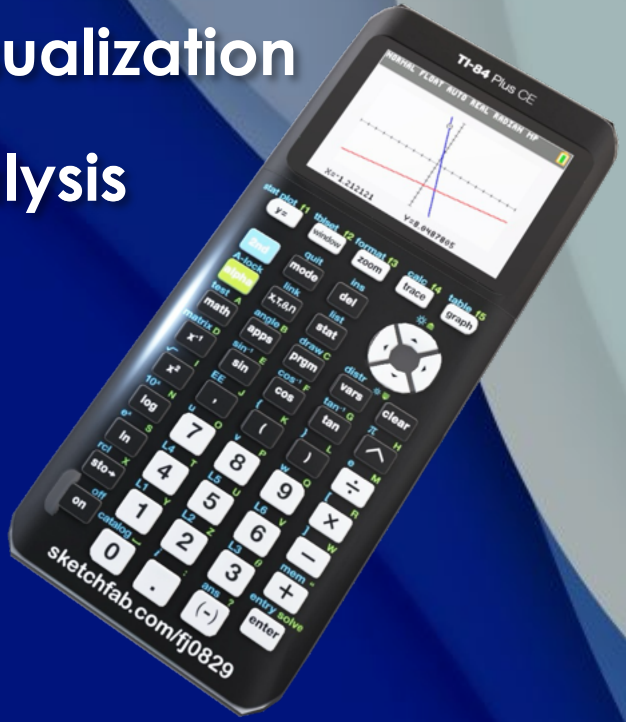
- Data analysis
- What is R?
- Fundamentals
- Data Structures
- Data Manipulation

Workshop 2:

- Control Sequences
- Data Wrangling
- Transformations

Workshop 3:

- working with dates and times
- advanced visualization
- statistical analysis



sketchfab.com/fj0829

Data analysis

- Imagine you are a detective , but instead of solving crimes , you are uncovering the story hidden within numbers and facts.

- This is what data analysis is about- finding patterns, answering questions, and making informed decisions based on data .

- R is often used to perform data analysis

What is Data?

- Data are pieces of information that can be collected and analyzed

- Forms of data:
  - Numbers, words , images, sounds
  - anything that can be measured

- When you check the weather, read reviews to decide on a movie, or compare prices before making a purchase, you're using data.

Types of Data

Two main types of data:

- **Qualitative**: Describe qualities or characteristics
  - color of a car
  - the flavor of a cake

- **Quantitative :** Data is numerical.
  - Number of students in a class
  - Number of participants in a study
  - temperature on a summer day

What is a population?

- Population isn’t just a group of people. It is the entire set of subjects or items we are interested in studying. This could be :
  - all the trees in a forest
  - every book in a library
  - all residents in a city

What is a sample?

- Most times, it is impractical, or impossible to study an entire population.

- So, we take a sample, a smaller group selected from the population , which is manageable yet representative enough to draw conclusion about the whole

sketchfab.com/fj0829

Empiricism

- Empiricism is the principle that knowledge comes from experience and evidence.
- In data analysis, it means making conclusions based on what we can observe and measure rather than just theories or beliefs.

Empiricism

- Example
  - $\mu =$ PARAMETER
  - $\bar{x} =$ STATISTIC

1. Population of $N = 10$ people of different height (inches).
   - Heights: 60, 71, 76, 56, 52, 65, 49, 53, 67, 64
2. Calculate $\mu = \frac{\sum x_i}{N} = \frac{60+71+76+56+52+65+49+53+67+64}{10} = \frac{613}{10} = 61.3$
3. Sample of $n = 3$ (71, 56, 64), Sample of $n = 6$ (76, 65, 53, 56, 60, 52)
4. Calculate $\bar{x} = \frac{\sum x_i}{n} = \frac{71+56+64}{3} = \frac{191}{3} = 63.67$
5. Calculate $\bar{x} = \frac{\sum x_i}{n} = \frac{76+65+53+56+60+52}{6} = \frac{362}{6} = 60.33$

Operationalism

Making concepts measurable:

- Operationalism is turning a concept into a quantifiable term.
- For example, how do we measure 'health'? We operationalize it by looking at indicators like blood pressure, heart rate, and cholesterol levels.
- It's how we define concepts so we can measure them.

Variables

- Variables are any characteristics, numbers, or quantities that can be measured or controlled.

- In the question of 'health', variables could be age, weight, diet, or exercise frequency.

- Variables are the basic units of data we analyze.

Descriptive vs. Inferential Statistics

Descriptive statistics :

- summarize and organize data so it's easier to understand. They provide a quick glance at the data through **averages**, **percentages**, and **patterns** without drawing conclusions about what the data means.

Inferential Statistics:

- While descriptive statistics give us the 'what' of the data, inferential statistics tell us the 'why'. They allow us to make predictions and inferences about a population based on the sample data we've collected.

Descriptive example

Imagine a teacher has the final grades for a class of 30 students. The teacher could use descriptive statistics to:

- Calculate the average grade (mean).
- Determine the grade smack in the middle (median).
- Identify the most frequently occurring grade (mode).
- Calculate the standard deviation to see how much grades vary.
- Create a histogram to visually represent the distribution of grades.
- Find the highest and lowest grade (range).

Inference example

Let's say a health researcher wants to estimate the average blood pressure of all adults in a city. It would be impractical to measure the blood pressure of every adult, so the researcher collects data from a sample of 200 adults.

Using inferential statistics, the researcher could:

- Use the sample mean blood pressure as an estimate of the population mean blood pressure.
- Create a 95% confidence interval to express the uncertainty of the estimate.
- Test a hypothesis, such as whether the mean blood pressure differs between males and females.
- Use regression analysis to predict blood pressure based on factors like age, weight, and exercise habits.

What is R?

- R is a programming language
- Open-source software via the GNU General Public License
- Widely used for statistical computing, data analysis, and graphics.
- It was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand.
- R is particularly popular among statisticians, data scientists, and researchers for its extensive statistical and graphical capabilities.

Where is R used?

Google

Microsoft

LLOYD'S
LLOYD'S OF LONDON

Azure Machine Learning

XBOX LIVE

IBM

intertrend

sketchfab.com/fj0829

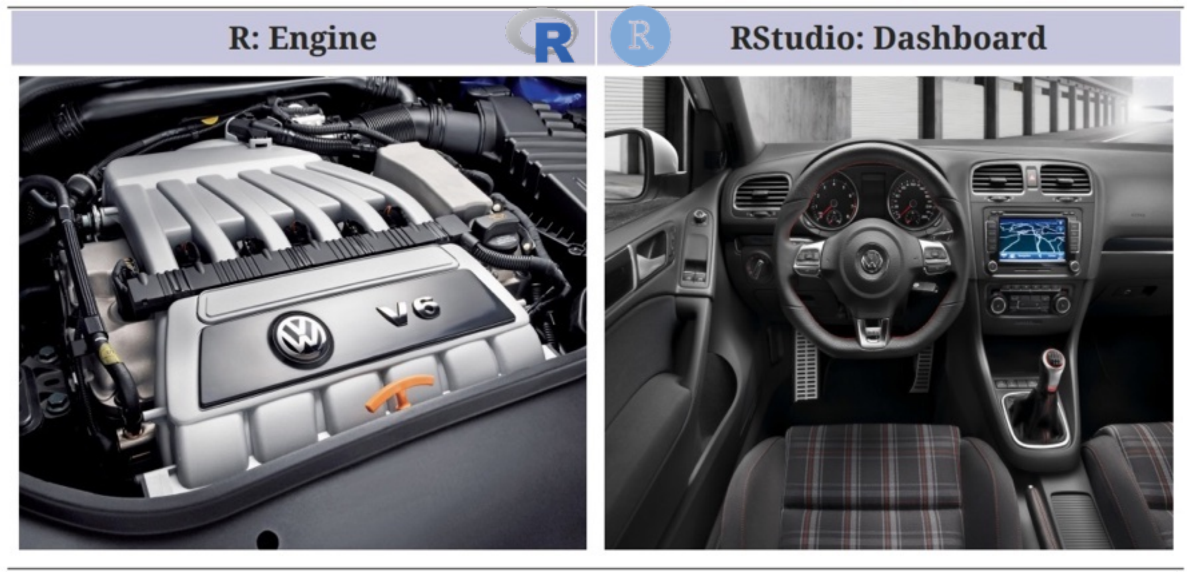
R & R packages

- R is mostly used in academia
- Interpreted language (Python is also interpreted)
- Different than excel , STATA, SAS.
- You interact with R via code (text instructions)
- R packages are pre-made instructions ready for you to use

sketchfab.com/fj0829

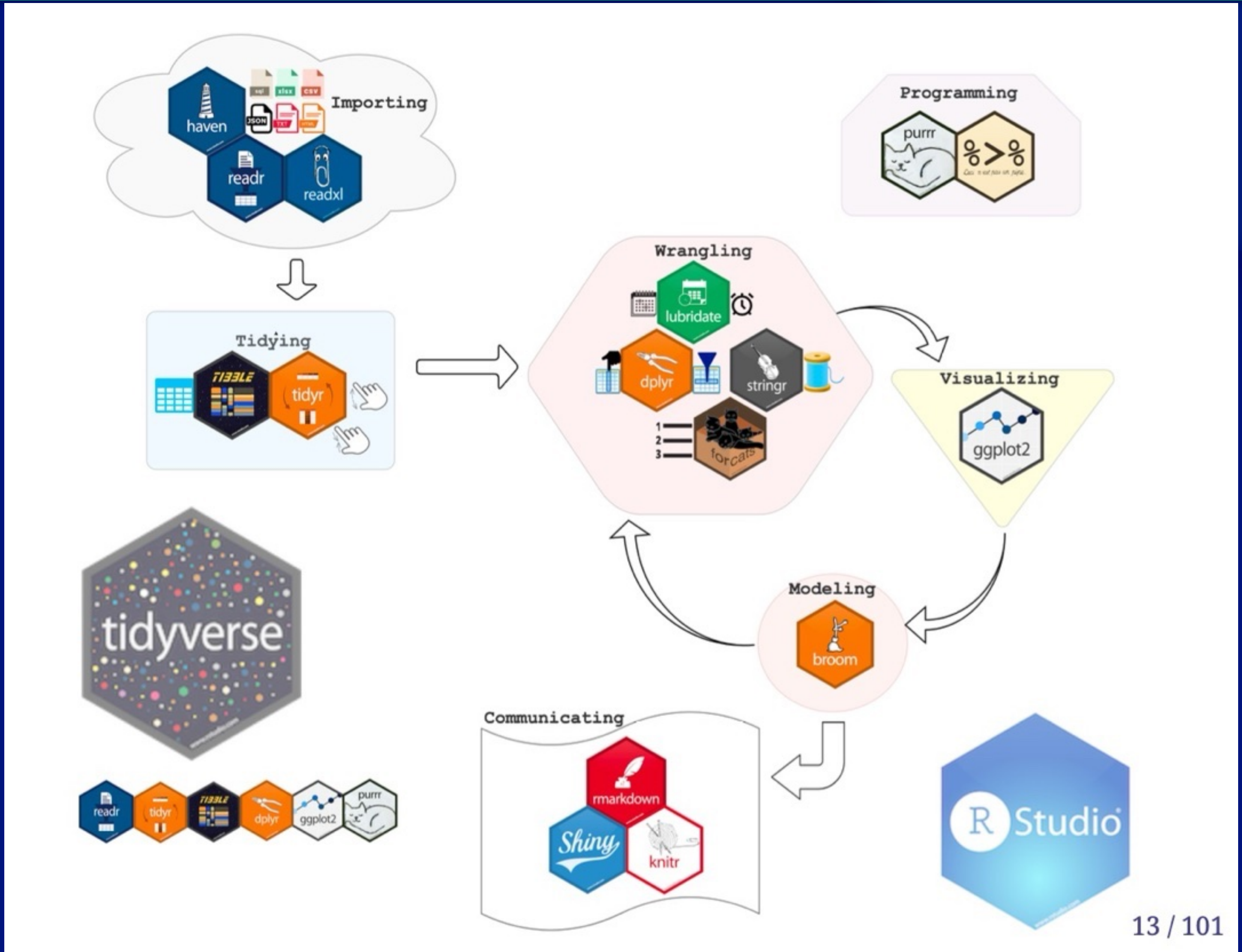
R vs RStudio

- R is like a car’s engine
- RStudio is like a car's dashboard, an integrated development environment (IDE) for R.



https://github.com/katrienantonio/workshop-R/

R universe



13 / 101

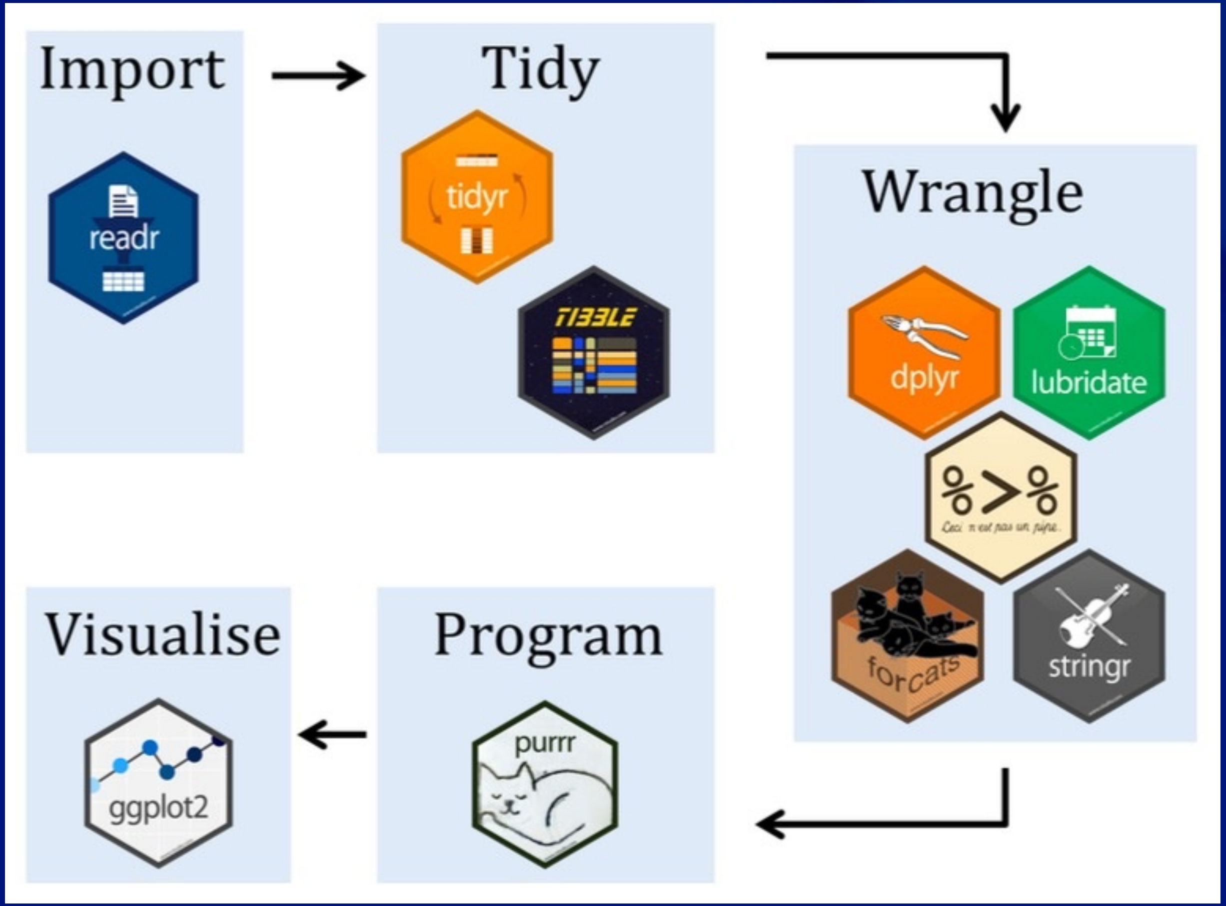
sketchfab.com/fj0829

https://github.com/katrienantonio/workshop-R/

Data Science workflow

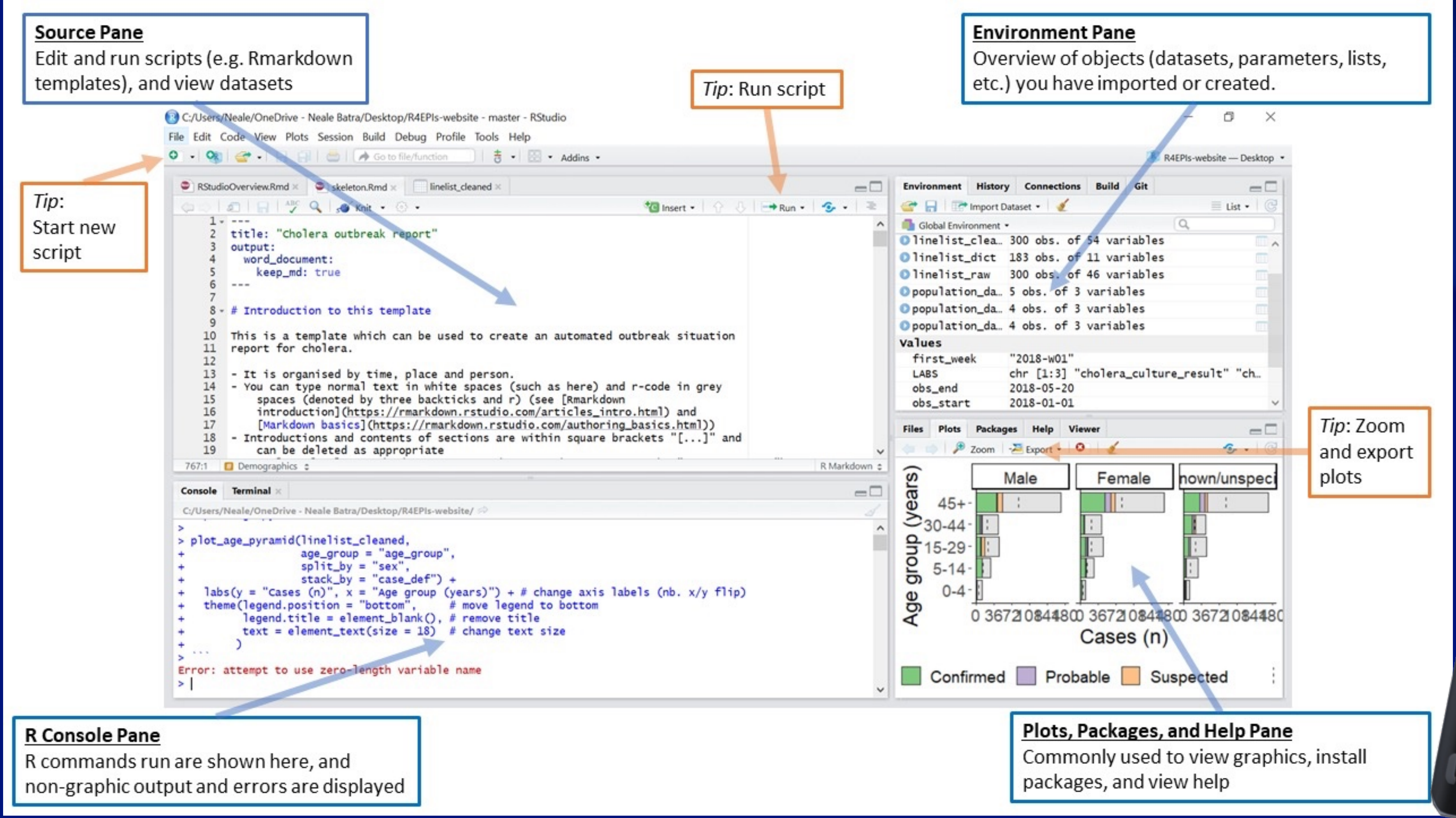
The **tidyverse** is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

https://docs.posit.co/resources/tidyverse/



https://github.com/katrienantonio/workshop-R/

RStudio Layout



Source Pane
Edit and run scripts (e.g. Rmarkdown templates), and view datasets

Tip: Run script

Environment Pane
Overview of objects (datasets, parameters, lists, etc.) you have imported or created.

Tip: Start new script

Tip: Zoom and export plots

R Console Pane
R commands run are shown here, and non-graphic output and errors are displayed

Plots, Packages, and Help Pane
Commonly used to view graphics, install packages, and view help

```
---
title: "Cholera outbreak report"
output:
  word_document:
    keep_md: true
---

# Introduction to this template

This is a template which can be used to create an automated outbreak situation
report for cholera.

- It is organised by time, place and person.
- You can type normal text in white spaces (such as here) and r-code in grey
    spaces (denoted by three backticks and r) (see [Rmarkdown
    introduction](https://rmarkdown.rstudio.com/articles_intro.html) and
    [Markdown basics](https://rmarkdown.rstudio.com/authoring_basics.html))
- Introductions and contents of sections are within square brackets "[...]" and
    can be deleted as appropriate
```

```
C:/Users/Neale/OneDrive - Neale Batra/Desktop/R4EPIs-website/
>
> plot_age_pyramid(linelist_cleaned,
+                  age_group = "age_group",
+                  split_by = "sex",
+                  stack_by = "case_def") +
+   labs(y = "Cases (n)", x = "Age group (years)") + # change axis labels (nb. x/y flip)
+   theme(legend.position = "bottom",      # move legend to bottom
+         legend.title = element_blank(), # remove title
+         text = element_text(size = 18)  # change text size
+        )
+  ...
>
Error: attempt to use zero-length variable name
> |
```

```
linelist_clea… 300 obs. of 54 variables
linelist_dict  183 obs. of 11 variables
linelist_raw   300 obs. of 46 variables
population_da… 5 obs. of 3 variables
population_da… 4 obs. of 3 variables
population_da… 4 obs. of 3 variables
Values
  first_week   "2018-W01"
  LABS         chr [1:3] "cholera_culture_result" "ch…
  obs_end      2018-05-20
  obs_start    2018-01-01
```

sketchfab.com/fj0829

https://sahirbhatnagar.com/EPIB607/basics.html

Website

tinyurl.com/R-WORKSHOP-JJAY

sketchfab.com/fj0829

Questions?

- To Access more materials and resources please visit the website on the QR code, or at **https://dilancaro.github.io/R-workshop-John-Jay/**
- Thank you