Introduction to R (TAU Workshop)

First session

Adi Sarid
Tel-Aviv University
Updated: 2020-12-07

Who am I?

- Adi Sarid
- Professional: Market Research, Data Scientist, Operations Research, R Educator
- Academia: PhD in operations research from TAU, the industrial engineering department
- Teaching the introduction to statistics and R course ("Data Analysis" 0560.1823) in the digital sciences for hi-tech program in TAU
- Software: R, Python
- Personal: lives in Netanya, father of three

Who is this workshop for?

- Since this workshop is given online, anyone is welcome, but specifically:
- This workshop is for students in the digital science for hi-tech program in TAU
- No prior knowledge in `R` programming language
- Studying introduction to statistics in the management school (can't attend the built-in intro statistics course of the digital sciences program)

Workshop Goals

- Students will acquire `R` skills
- Learn how `R` can be used for preparing data, visualizing data, and modeling data
- We will assume that you already know the theory behind statistics, and focus on practical (and fun!) stuff
- **By the end of the workshop students will be able to take a data set and analyze it in meaningful ways, drawing conclusions, and presenting the results.**

What to expect today?

- Today you will build your first plot in `R`!
- You will visualize google mobility trends in Israel, and discuss how they were affected by Covid-19
- You will see how data is structured (and discuss tidy data)
- You will discuss the data origin and see problems with it
- We will see the RStudio IDE and demonstrate base-R syntax

Technical details

Our workshop is going to be comprised of:

- Lectures: background and explanations
- Demonstrations: live coding and analysis of data sets
- Labs: these are either tutorials as we will do today, or coding in your computer in R
- Optional home assignments: tasks I will give each week (which I will solve in the following session)

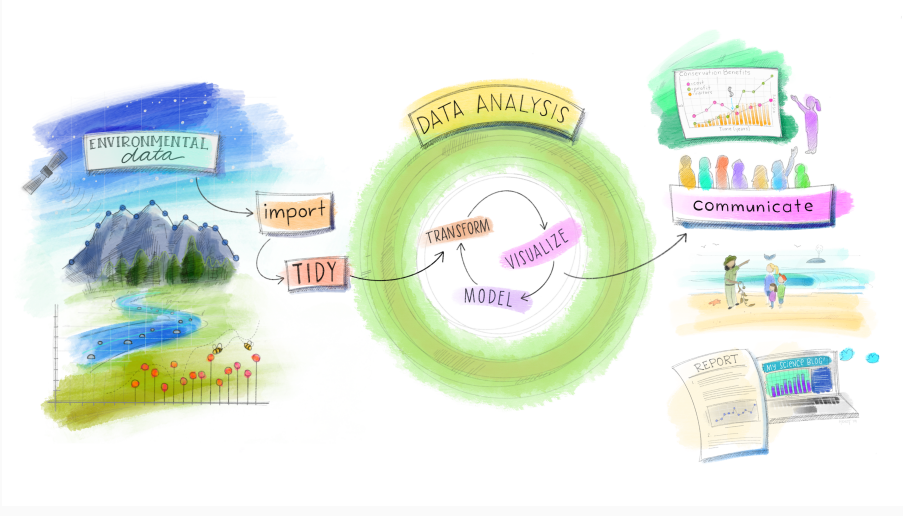
Tech stack / sources

- All the code and materials are available on github: https://bit.ly/tau-r-workshop
- You will need R and RStudio Desktop which are both open source and free to download and use
- The `tidyverse` package (`install.packages("tidyverse")`)
- Syllabus, books, youtube channels, and other sources: see github repo README.md under reading materials.

Covid19 mobility trends

- For this lab we're going to split to breakout rooms
- Each group will solve the exercise here: sarid.shinyapps.io/covid19_mobility
- We will give 20 minutes for this exercise, but I will circulate among the rooms and see if you need more/less time
- Upon completion, we will solve the exercise together

Some context: the data science workflow



Source: Illustrations by Allison Horst

Back to the basics

Switching to some live coding session:

- We will demonstrate base R syntax
- We will familiarize ourselves with the RStudio IDE

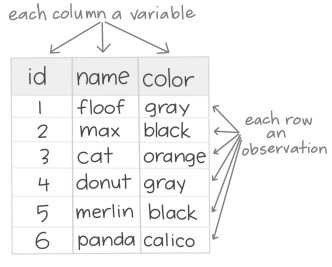
Tidy data: a friendly definition

"TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."

–HADLEY WICKHAM

In tidy data:

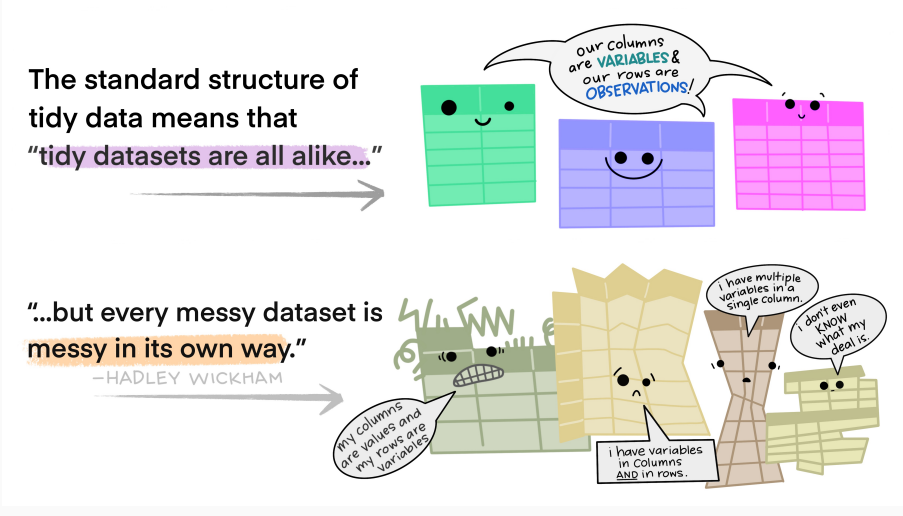
- each variable forms a column
- each observation forms a row
- each cell is a single measurement



Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10

Illustrations from the Openscapes blog Tidy Data for reproducibility, efficiency, and collaboration by Julia Lowndes and Allison Horst.

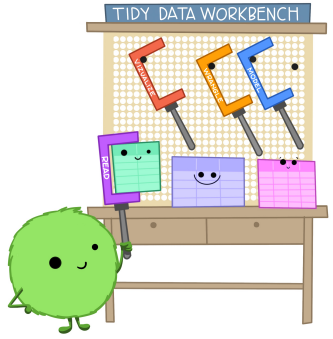
Tidy data: tidy versus untidy



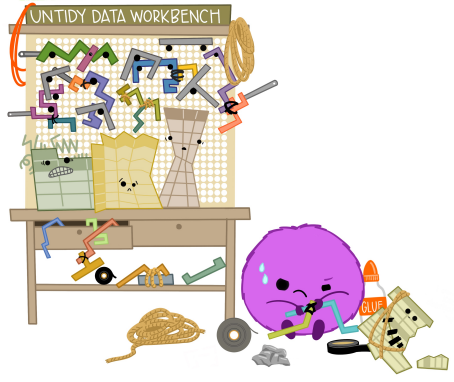
Excel users have a tendency to get data to be untidy: merged cells, colored cells, aggregated cells, skipping rows, hidden columns, formulas, pivot tables, etc...

Tidy data: a consistent set of tools

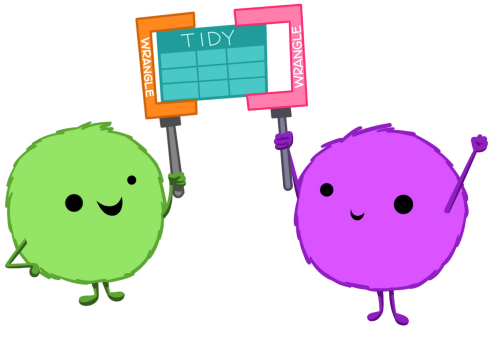
When working with tidy data, we can use the same tools in similar ways for different datasets...



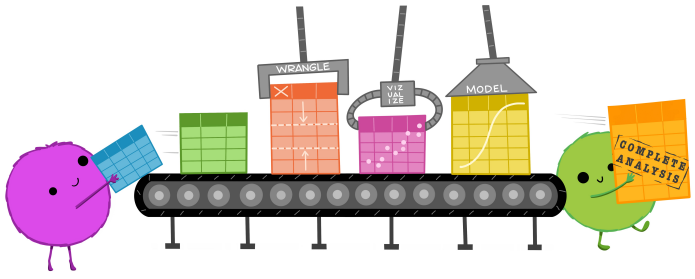
...but working with untidy data often means reinventing the wheel with one-time approaches that are hard to iterate or reuse.



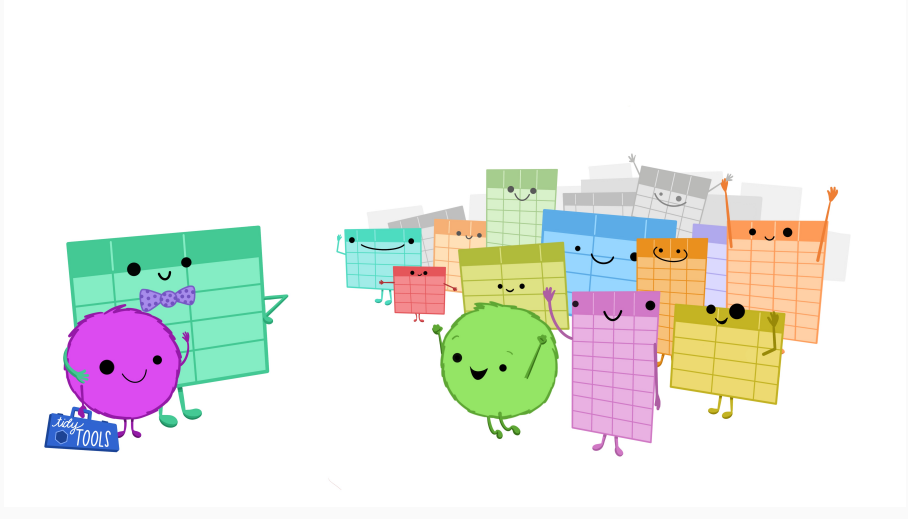
Tidy data: we all speak the same language



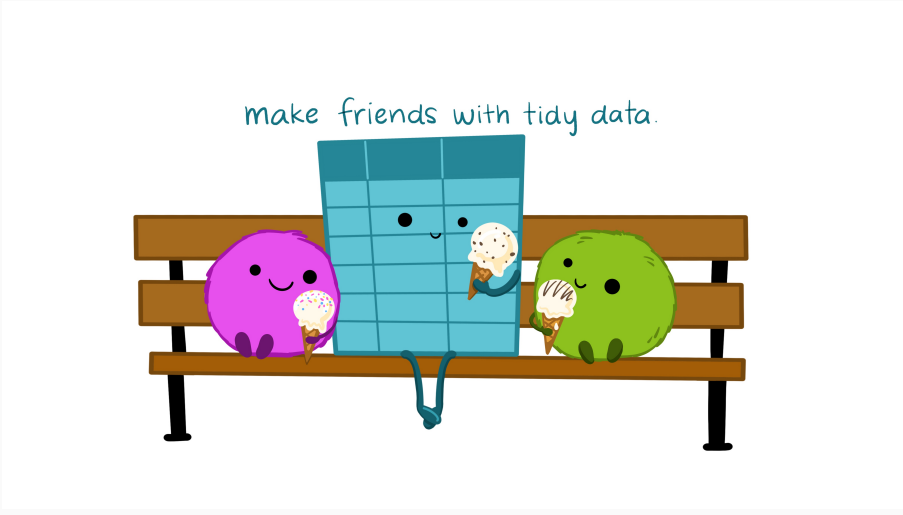
Tidy data: and can automate many tasks



Tidy data: tidy is good



Tidy data



Class exercise (in groups, if time permits)

The Fibonacci series is a series in which every element is the sum of the previous two elements, i.e.:

$$1, 1, 2, 3, 5, 8, 13, 21, 34, \ldots$$

Use the following code to build a loop that prints out the first 20 elements of the Fibonacci series:

```
# Fibonachi code exercise, fill in the blanks (where you see `?`)
total <- ?

element_i_minus1 <- 1
element_i_minus2 <- ?

for (? in 3:total){
  next_element <- ? + ?
  element_i_minus2 <- element_i_minus1
  ? <- next_element
  cat("\n", ?)
}
```

Wrap up

- We have seen how R can be used for visualizations and along the way learned important aspects such as the data science work flow and tidy data.
- You experienced a bit of coding and making visualizations on your own - don't worry we will get back to visualizations in our third session.
- We have seen RStudio IDE, and base R (assignments, functions, conditions, logical operators, special values).

Optional homework

Read about tidytuesday here. Watch a tidytuesday video here.

Next week we will talk about tidyverse and how to tidy and transform data.