Introduction to R (TAU Workshop) Data Transformation and Wrangling

Second session

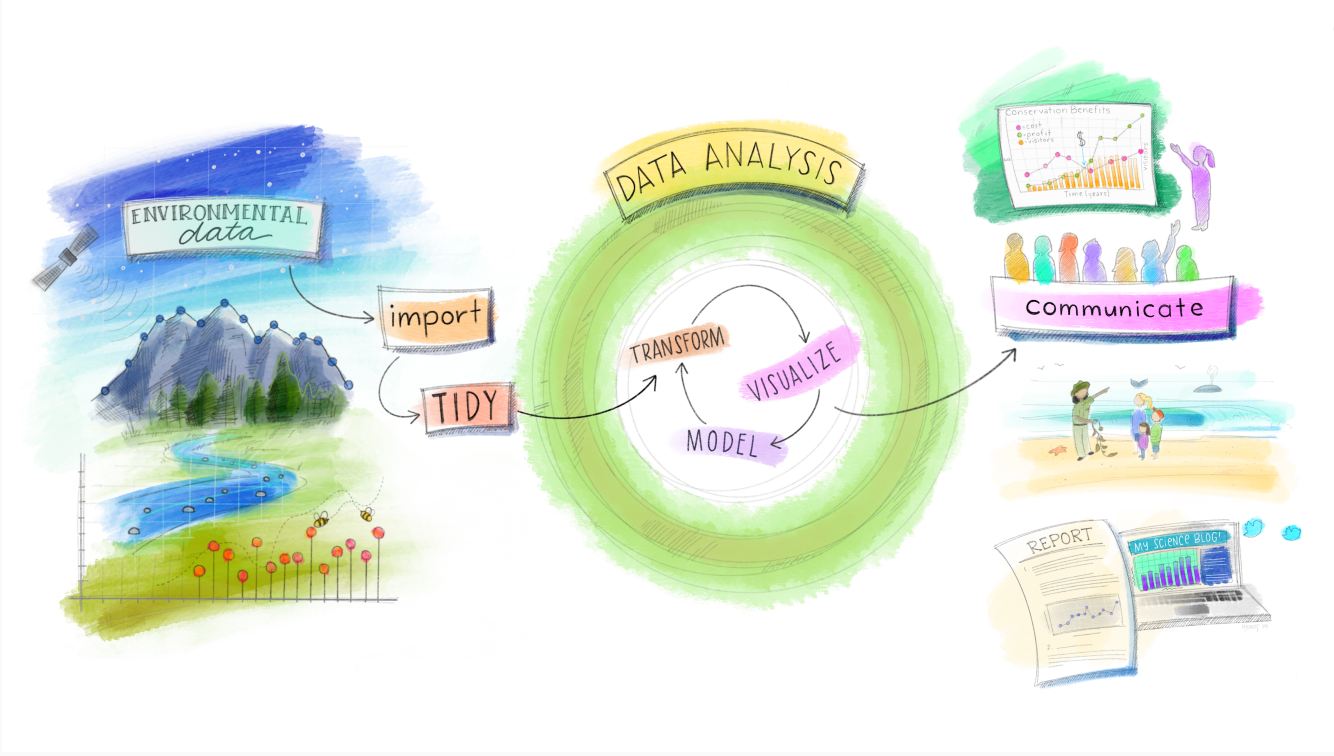
Adi Sarid
Tel-Aviv University
Updated: 2020-12-14

Recap from last week

In the previous session

- You built your first plot in `R`!
- You visualized google mobility trends in Israel, and discussed how they were affected by Covid-19
- You saw how data is structured and discussed tidy data
- We discussed the data science workflow
- You saw the RStudio IDE and base-R syntax (i.e., for loops, if clauses, functions, and more)

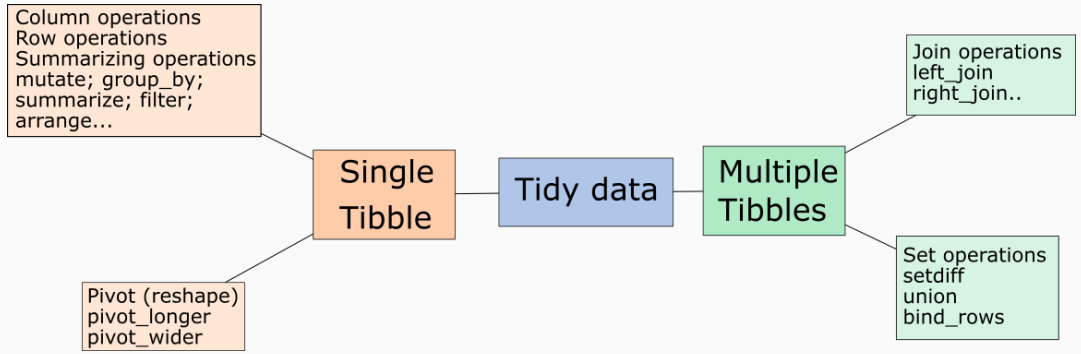
Today's focus: data transformations and wrangling



What we will do today

- Focus on tidyverse syntax
- Some quick tricks on getting to know your data
- Learn how to transform data sets so they will be easier to analyze:
  - Operations that affect rows (e.g., filtering, arranging)
  - Operations that affect columns (e.g., selection, mutation)
  - Operations that summarize data (e.g., compute mean, sd, percentiles)
  - Changing the data representation (changing long to wide formats or vice verse)
  - Merging (joining) data sets

Conceptual map of basic tidyverse operations



Some quick tricks to get to know your data

Prerequisite: throughout the exercise we will use the `tidyverse` packages. To install them use `install.packages(tidyvers)` and then run `library(tiydverse)` in your script.

- The pipe is very useful `%>%`

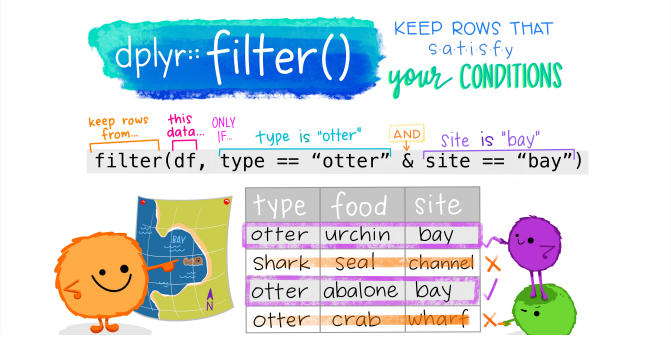
When you explore a dataset for the first time, use:

- `glimpse` to see what variables there are
- `head` (and `tail`) to see the first few first (and last) set of lines
- `View` to see the data in a table
- `count` to count a specific variable or combination of variables

Operations that affect rows (manipulate cases)

These are operations like sorting the rows according to a specific variable, filtering tables, sampling, or taking distinct values

- `filter` (remember logical operators?)
- `arrange`
- `sample_n` and `sample_frac`
- `slice`
- `distinct`



Operations that affect columns (manipulate variables)

These are operations that help you transform variables, create new variables, select specific variables

- `select` to select specific variables
- `mutate` to create or modify variables
- `mutate_*` and `across`

Operations that summarize data (summarize cases)

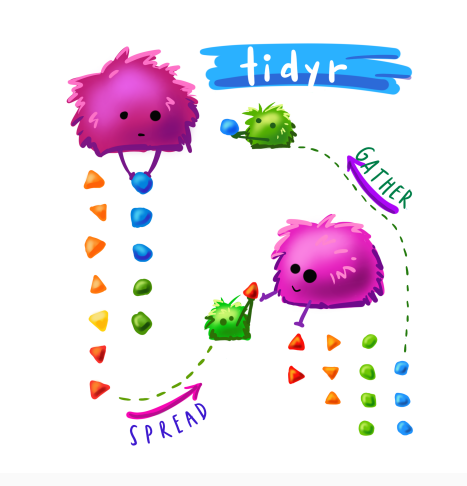
Apply summary functions to create a new table. To summarize data you need to combine `group_by`, `summarize`, and a summary function:

- `group_by` and `summarize`, followed by:
- `sum`, `n`, `mean`, `median`, `first`, `last`, `quantile`, `min`, `max`, `sd`

Changing the data representation (long vs. wide)

Very often you will be required to change the data representation, from wide to long format (or vice verse).

- This is called pivoting, and you can use `pivot_longer` or `pivot_wider`
- Especially useful when plotting with `ggplot2`, which we will discuss next week



Joining data sets (combine tables)

Sometimes we have multiple sources of data, and we need to join them, or add them to one another

- Joining = Combining variables, adding new columns using shared key variables (`left_join`, `right_join`, `full_join`)
- Binding = Combining cases, adding new rows from another source, which has the same variables (`bind_rows`)

Summarising exercise

In this exercise you will explore (transform and wrangle) Himalayan Climbing Expeditions.

- Expeditions that climbed in the Nepal Himalaya
- To complete this exercise you will need R, RStudio IDE.
- Download and follow the instructions here

Wrap up

- We have seen a lot of tidyverse basics for data transformations and wrangling
    - Rows affecting functions, column affecting functions, summarizing functions, pivoting
    - These are all the must-have elements for data analysis
    - You want to exercise these as much as you can if you want to become proficient in data science
- You experienced the various functions using the Himalayan Climbing Expeditions data set from the tidytuesday repository

Optional homework

Analyze the IKEA data set, and answer:

- How many different items appear in the data set?
- What is the most frequently appearing category?
- What is the most expensive category (on average? on median?)
- How many different items are sellable online?
- What is the average discount provided at the time of data collection?
- What is the average size of a chair?
- Create a new table with each row representing a product name, and each column representing a category. The table values should represent the number of available products with that name and product category.