BST 219
Core Principles of Data Science

Lecture 15: Advanced Data Wrangling continued
October 22, 2024

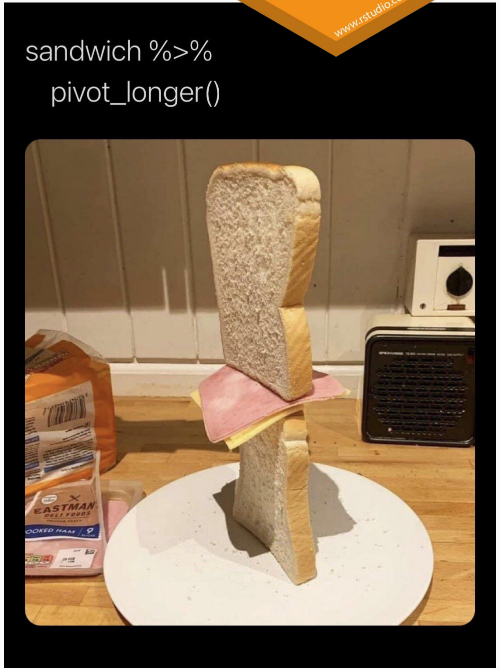
Recipe of the Day!

Turkish Cig Kofte

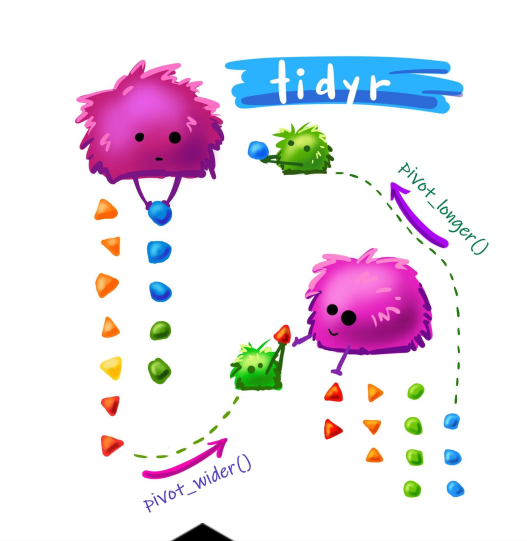
Emre says he'll help you with your homework if you help him make a castle with magnet tiles

Agenda

- Announcements
  - Lab this week!
  - Homework 2 is due 10/25
- Review mid-course evaluation survey results
- Chat about the midterm and final project
- Continue the advanced data wrangling module
  - Importing data
  - Reshaping data



```
sandwich %>%
  pivot_longer()
```



Mid-course evaluation survey results

What is going well to help learning

- The coding question of the day
- Lab problems
- Lecture and lab recordings
- Using RMarkdown files
- Breaking down why and how we code certain things
- Snacks

What could be done to improve learning

- More practice problems
- TFs up to date with content that has been presented in class
- Having a TF in lecture to help with individual coding issues
- Instead of a few long assignments, split the questions into several shorter assignments to speed up feedback
- Clarify deadlines for upcoming assignments
- Lab recordings with fewer tech issues
- More hints for certain HW questions because they can be a bit vague
- Providing answers to assignments and in-class examples

Midterm and Final Project

- **Midterm**
  - Will be given in November, final date will be announced ASAP (we need to finish the advanced wrangling module and maps)
  - Will be a take-home exam
  - Will be due a week after it's assigned
  - Will be submitted exactly like a homework assignment

- **Final Project**
  - Details are in the syllabus and on the course website
  - Need to form groups of 4-5 students
  - Need to submit a project preference form with the names of all group members
    - 1 form per group
    - Due **November 12**
  - A TF will be assigned to each group to help guide the analysis plan
  - All materials due December 16

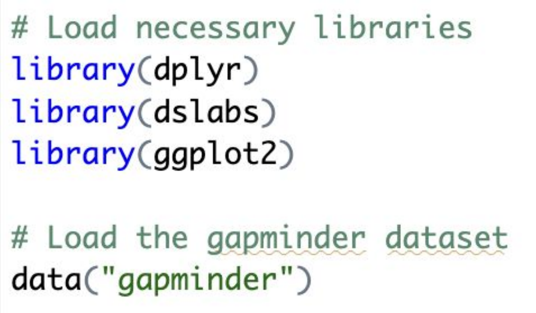
Coding Question of the Day!

Using the **`gapminder`** dataset, create a categorical life expectancy variable and add it to the gapminder dataset. Call this variable **`life_expectancy_category`** and use the **`case_when`** function to create it. Life expectancy below 50 should be labeled "**Low**", life expectancy greater than or equal to 50 and less than 75 should be labeled "**Medium**", and life expectancy greater than or equal to 75 should be labeled "**High**".

Now, filter the dataset to only include observations from the year 1999, and create a bar chart of the 3 life expectancy categories.

Bonus challenge: Reorder the bars so that they are in the order Low, Medium, High. You can use whatever function you prefer or find online that works.

Make sure to run this code first



```
# Load necessary libraries
library(dplyr)
library(dslabs)
library(ggplot2)

# Load the gapminder dataset
data("gapminder")
```