BST 219
Core Principles of Data Science

Lecture 9: R Basics Continued
October 1, 2024

Recipe of the Day!

Butternut Squash Bisque

Happy October!

Agenda

- Announcements
  - No lab this week!
  - If you had issues committing and pushing your homework, please come to office hours or set up a time to meet with me or one of the TFs

- Start the visualization module!

Coding Question of the Day!

Using the gapminder dataset, for the year 2000, add a column to the dataset called `high_life_exp` that is equal to 1 if the life expectancy is above 70 years of age and 0 otherwise.

How many countries had a life expectancy above 70 years in 2000? What percentage of countries included in the dataset is this?

Hint: it might be useful to use the **ifelse** function

Make sure to run this code first

```
library(dslabs)
library(dplyr)
data(gapminder)
```

Where we are in the data science pipeline

(But we'll come back to this with more advanced techniques!)

Importing (loading) the data

Processing (cleaning, wrangling) the data

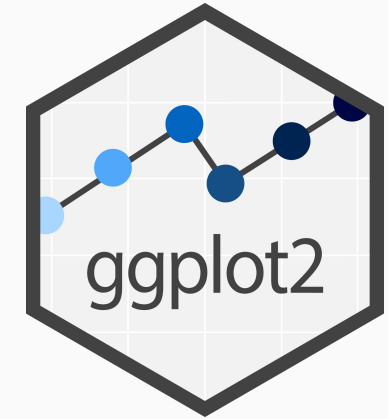
Visualizing and summarizing the data

Building models (statistical and ML)

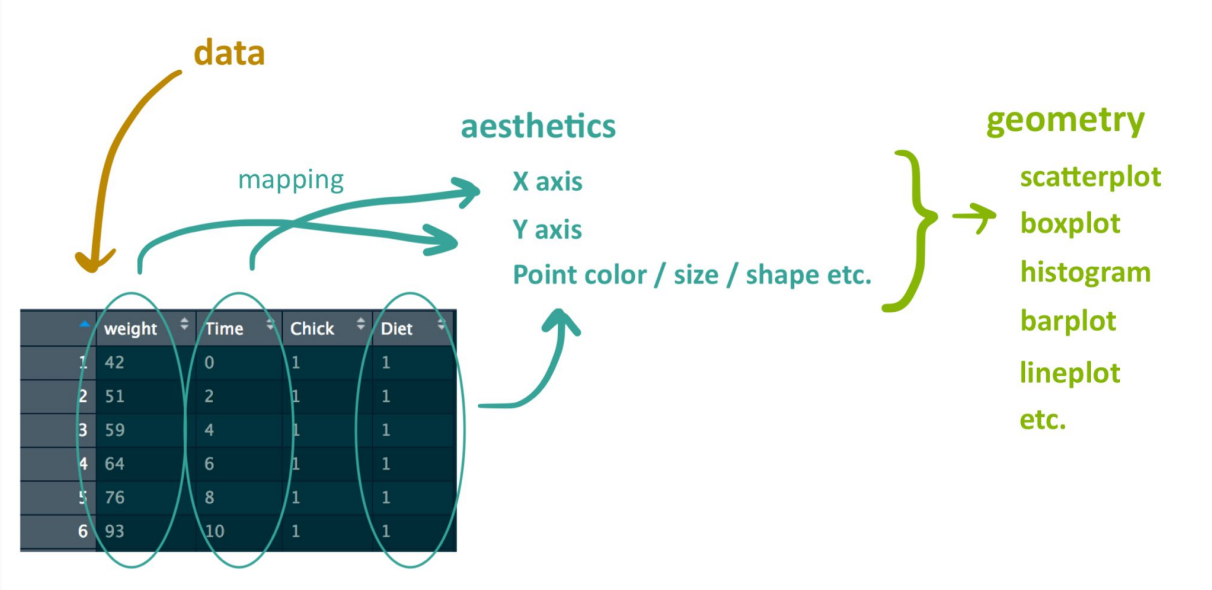
Interpretation and communication of results

ggplot2

- Plots are built using layers
- Each layer is usually 1 line of code
- Plots can be basic/simple, or complex
- Plots can be an efficient way to convey a lot of information
- Each plot should be easy to read/interpret
  - It is your job to make the plot as easy to interpret as possible for the reader or audience



ggplot2

- The “gg” part of “ggplot2” stands for the ***grammar*** of ***graphics***. Just like sentences are composed of various parts of speech (e.g., nouns, verbs, adjectives) that are arranged using a grammatical structure, ggplot2 allows us to create figures using a standardized syntax.

- 3 main components of a plot
  - Data
  - Aesthetics
  - Geometry