[COGS 9] Discussion Reading 3, Visualizations, Matplotlib

Reading Quiz 3 due on 10th Feb (Fri)

Assignment 1 due on 13th Feb (Mon)

Final Project discussions next Wednesday

Programming for Data Science

Why and How to Program?

- Tips on how to learn to code and resources

Different roles in the industry

- Software Engineer, Software Developer, Computer Scientist
- Data Scientist, Machine Learning Engineer, Data Analyst

R (in Academia), Python, SQL

- Imperative vs Declarative

Version Control (Git) -> Github

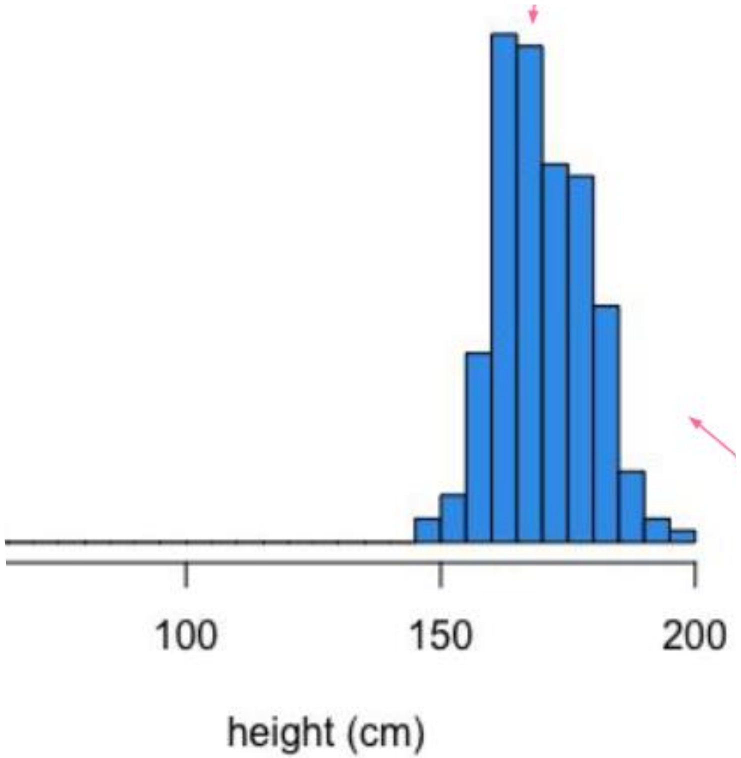
- GOLD MINE - https://education.github.com/pack/offers

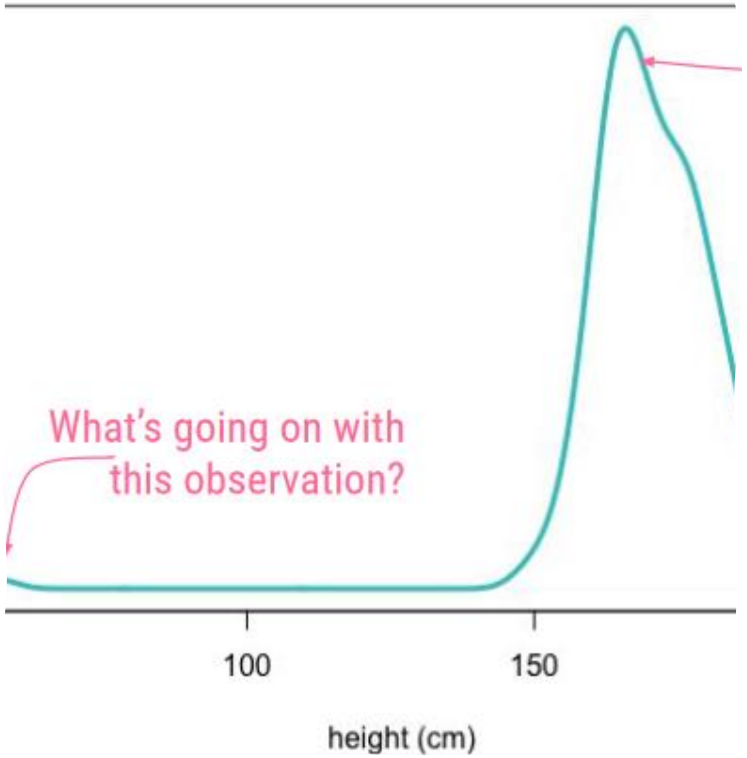
Python Data Stack

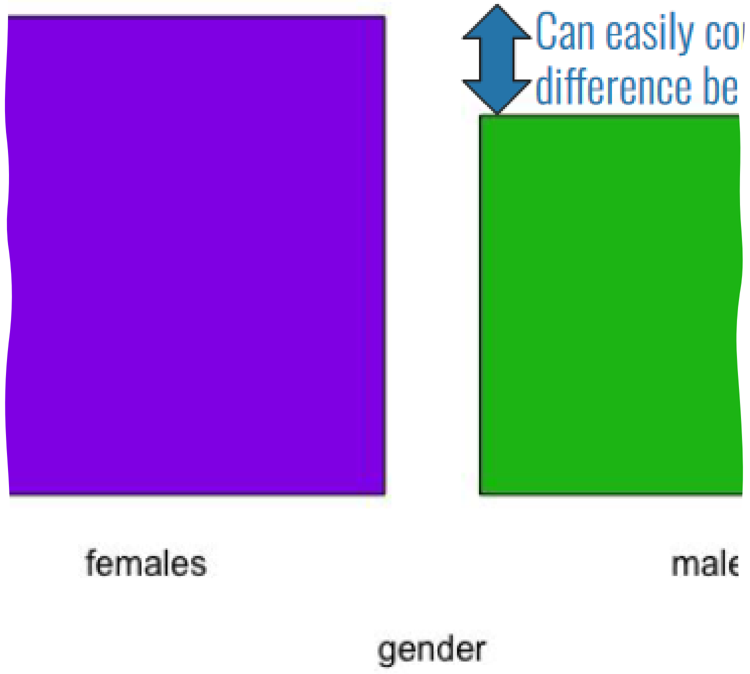
- Anaconda, Miniconda, Numpy, Pandas, Matplotlib, Seaborn, Scipy, SKLearn, PyTorch

Data Visualization

What’s the difference between a histogram, densityplot and barplot?

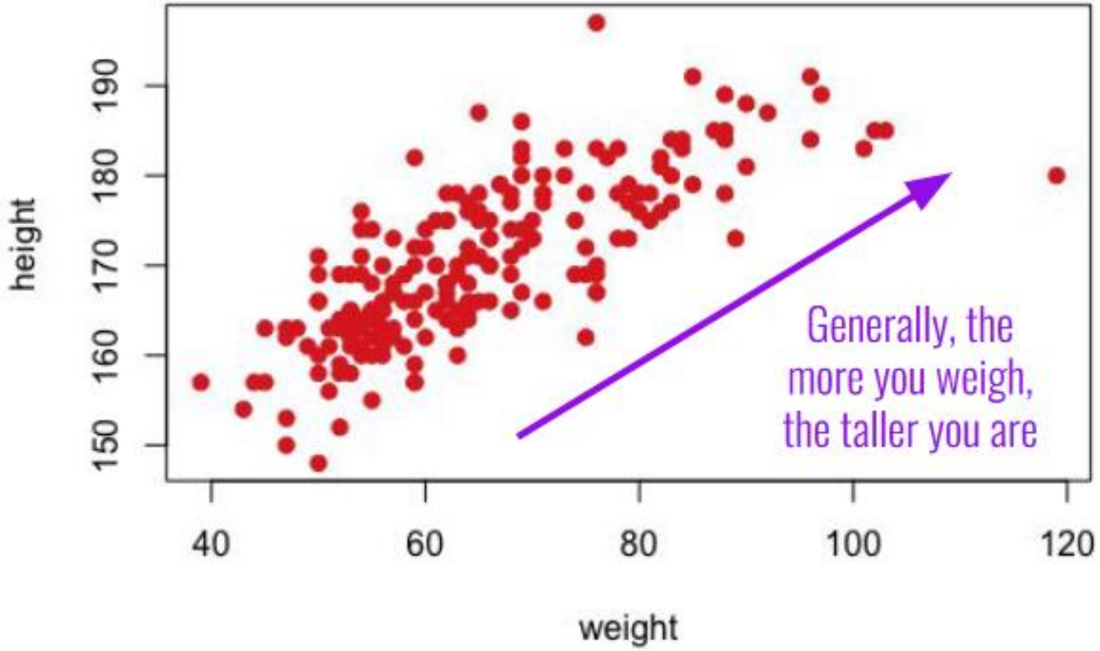


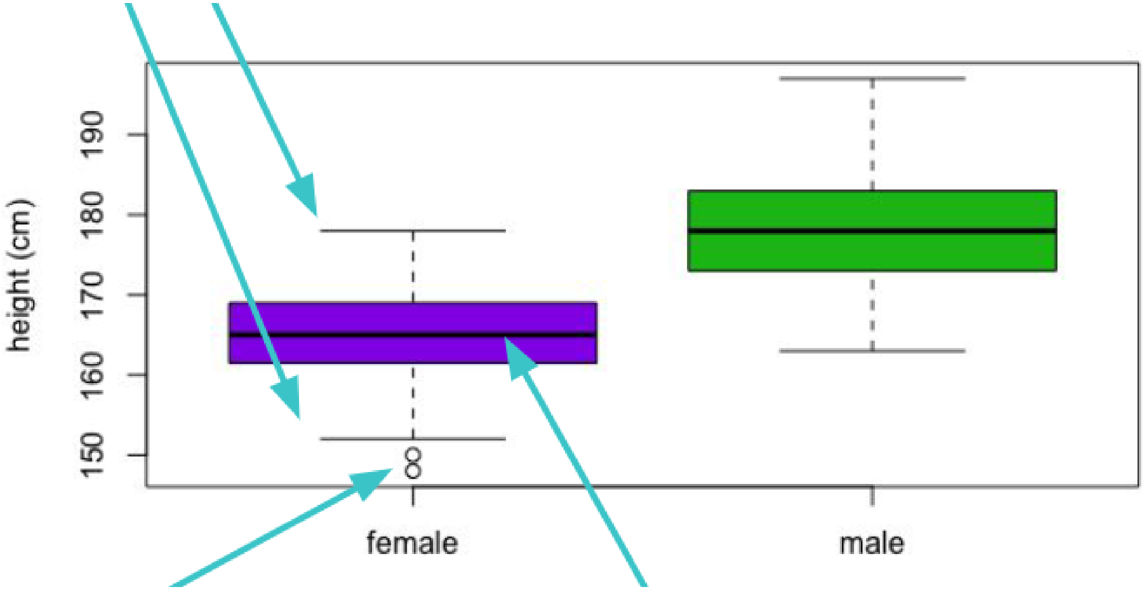




Data Visualization

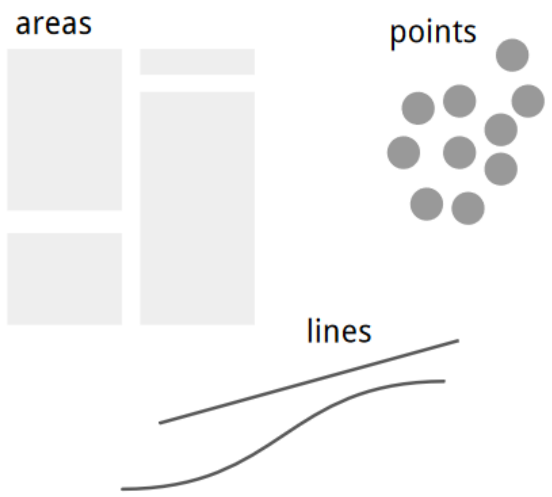
When to use a scatterplot and a boxplot?



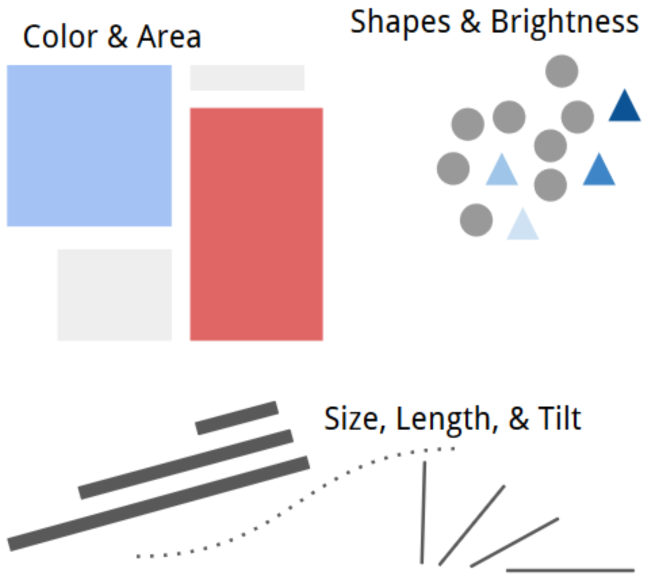


Data Visualization

Marks and channels



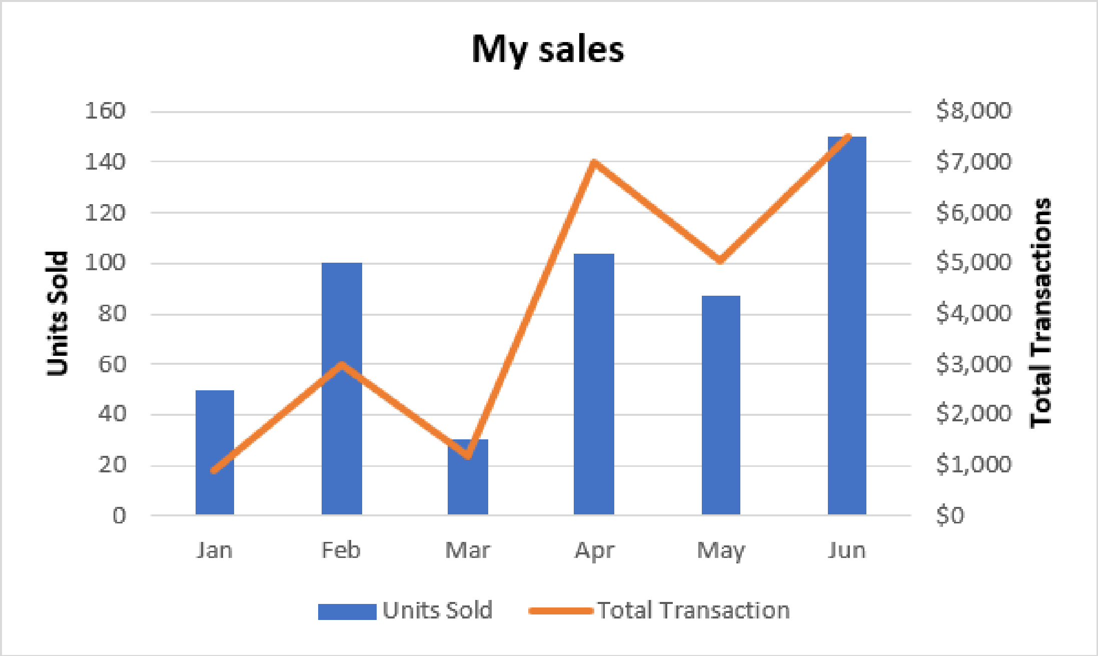
Marks: Geometric Primitives

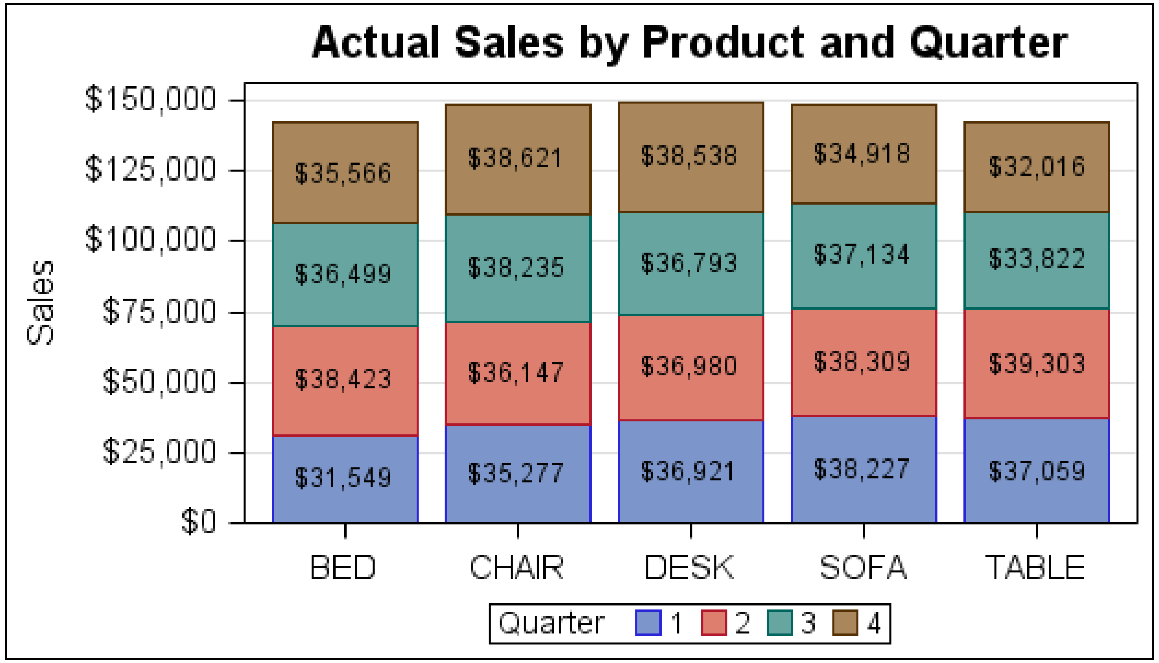


Channel: way to control the appearance of a Mark

Data Visualization

- Express: The visual shall express all of, and only, the information provided by the data's attributes. The visual should not add anything to/remove from the data.
- Effect: How important an attribute is, must match the salience of the channel. Greater importance = greater salience, or more noticeable





Data Visualization

Checklist when creating graphs

- Consideration for Colour-blindness
- Label the axes
- Ensure that the data is correct
- Ensure that the graphic represents the data
- Make the comparison easy on readers
- Ensure that the y-axis starts at 0 (What about x-axis?)
- Choose best visual
- Keep it Simple Stupid

Data Visualization

Checklist when creating tables

- Have a top to bottom comparison
- Logical row ordering
- Logical column ordering
- Limit number of rows and columns
- Informative headers
- Fix significant digits
- Format table

Descriptive Analysis

- Suppress some of the truth so that humans can understand easily
- Size, shape, missingness, central tendency, variability
- Size: Number of variables and observations
- Shape: Distribution of the variables (Uniform, bimodal, Normal/Gaussian/Bell-shaped, left & right skewed, random)
- Missingness: How much data is missing?
- Central Tendency: Mean, median, mode
- Variability: Variance, Standard Deviation, Range

Matplotlib Demo

https://matplotlib.org/stable/tutorials/index.html

https://matplotlib.org/stable/gallery/index.html