DATA VISUALIZATION USING PYTHON

D-VELOP WORKSHOP SERIES – Summer 2021

Trevor Bonjour

PURDUE UNIVERSITY | Libraries and School of Information Studies

D-VELOP WORKSHOP SERIES – Summer 2021

Jun 9	• Data Visualization: ggplot2
Jun 16	• Data Visualization using Python: Matplotlib and Seaborn
Jun 23	• Exploratory Data Analysis in R
July 7	• Data Visualization using Python: Bokeh (Interactive Plots)
July 14	• Exploring and Visualizing Time Series Data
July 21	• Data Visualization: introduction to Tableau

PURDUE UNIVERSITY | Libraries and School of Information Studies

What will we cover today?

- Motivation
- Useful Python Libraries
- Types of Plots
- Learn by Doing

Visualization Objectives

- Record information
- Analyze data to support reasoning
- Confirm hypotheses
- Communicate ideas to others

Why Visualize

To record information

Why Visualize

To point out interesting things

MTHIVLWYADCEQGHKILKMTWYN
ARDCAIREQGHLVKMFPSTWYARN
GFPSVCEILQGKMFPSNDRCEQDIFP
SGHLMFHKMVPSTWYACEQTWRN

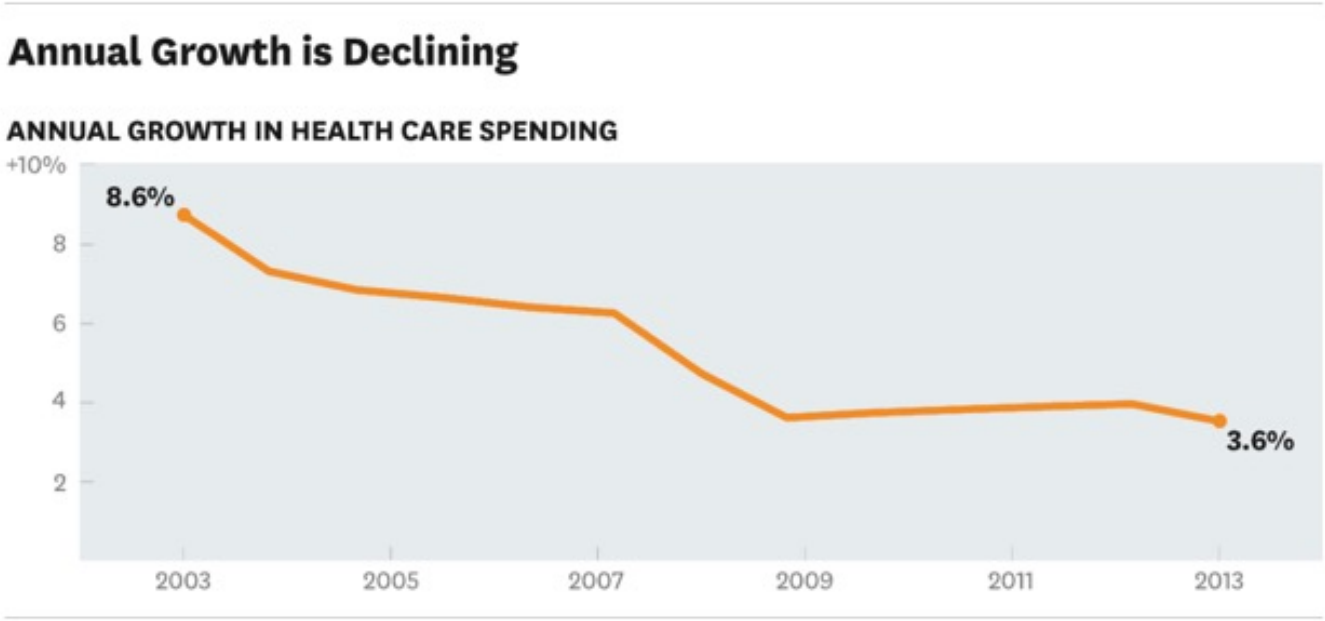
Why Visualize

To point out interesting things

MTHIVLWYADCEQGHKILKMTWYN
ARDCAIREQGHLVKMFPSTWYARN
GFPSVCEILQGKMFPSNDRCEQDIFP
SGHLMFHKMVPSTWYACEQTWRN

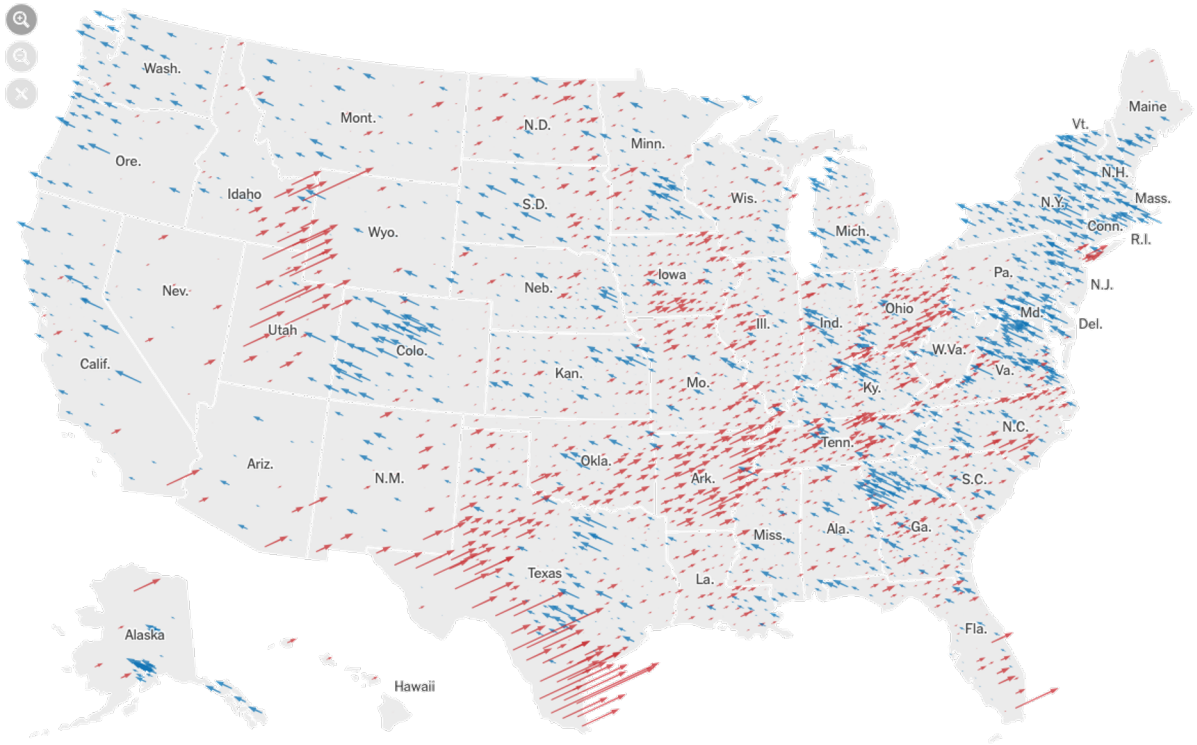
Why Visualize

To communicate information



Why Visualize

To analyze data



2020 US Elections (NYTimes)

Useful Python Libraries

NumPy

pandas

matplotlib

seaborn

bokeh

NumPy

- Fundamental package for scientific computing
- Exceptionally fast – written in C
- Main data structure:
  - *ndarray* : n-dimensional arrays of homogeneous data types
- Data manipulation ≈ NumPy array manipulation
- Used in other libraries - Matplotlib, pandas, scikit- learn

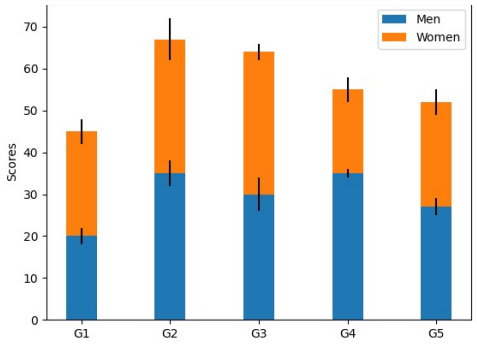
Link: NumPy for MATLAB USERS

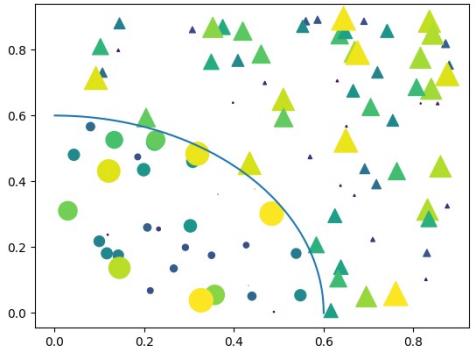
Pandas

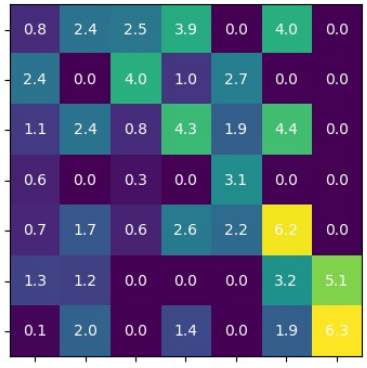
- Fundamental tool for handling and analyzing input data
- Particularly suited for tabular data
- Implements powerful data operations
- Main data structures:
  - *DataFrame*: A table with rows and columns
  - *Series*: A single column

Matplotlib

- Used for basic plotting
- Highly customizable
- Works with NumPy and pandas

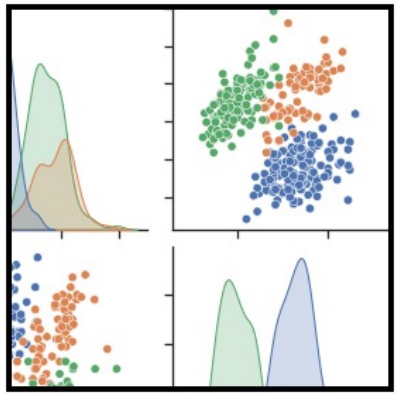


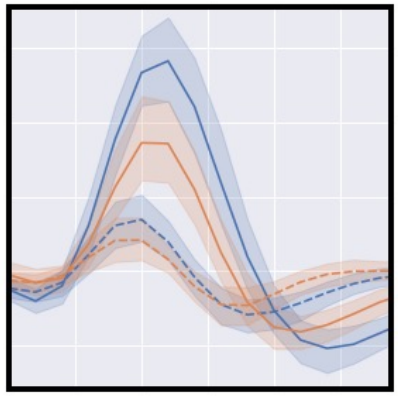


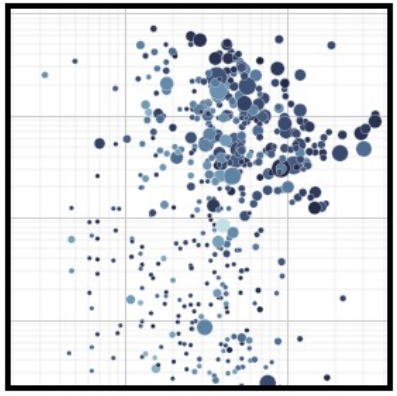


Seaborn

- Used for statistical data visualization
- Uses fewer syntax with good default themes
- Integrated to work great with pandas data-frame
- Uses Matplotlib under the hood

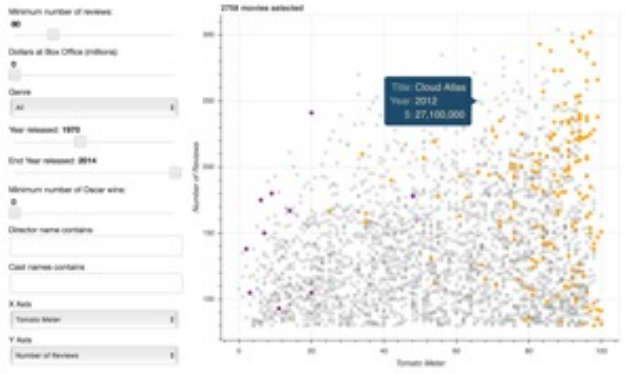


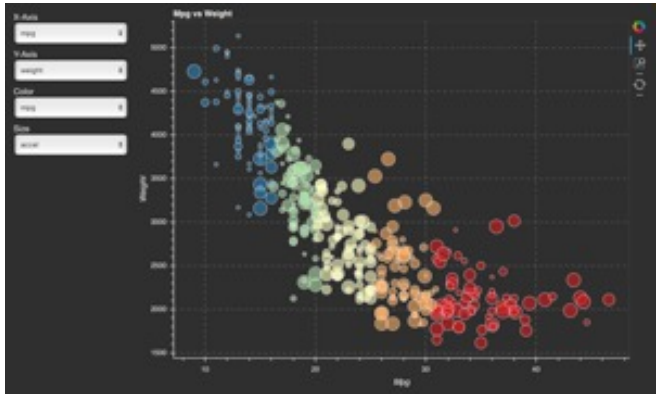




Bokeh

- Used for interactive visualization
- Requires modern web browsers
- Integrates with JavaScript



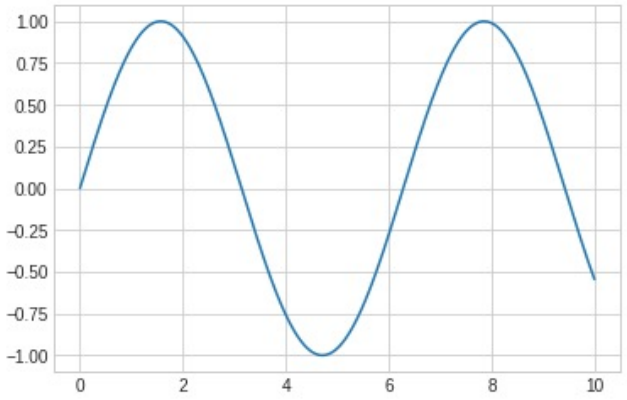


Types of Plots

- Line plots
- Bar plots
- Scatter plots
- Box plots
- Histograms

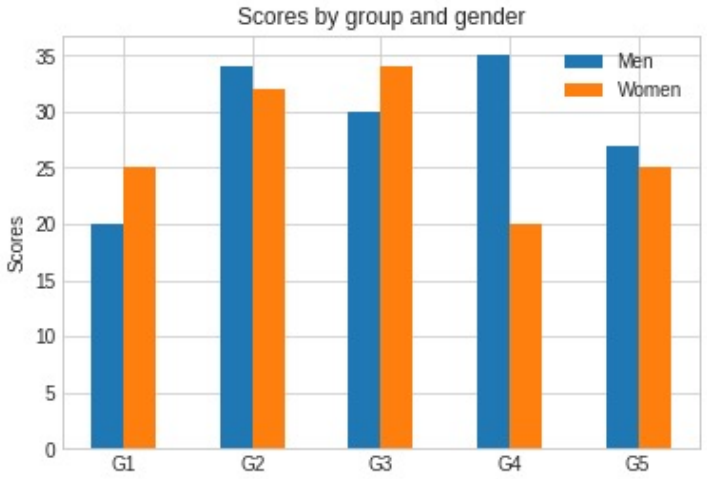
Line plots

- Used for numeric data
- Used to show trends
- Compare two or more different variables over time
- Could be used to make predictions



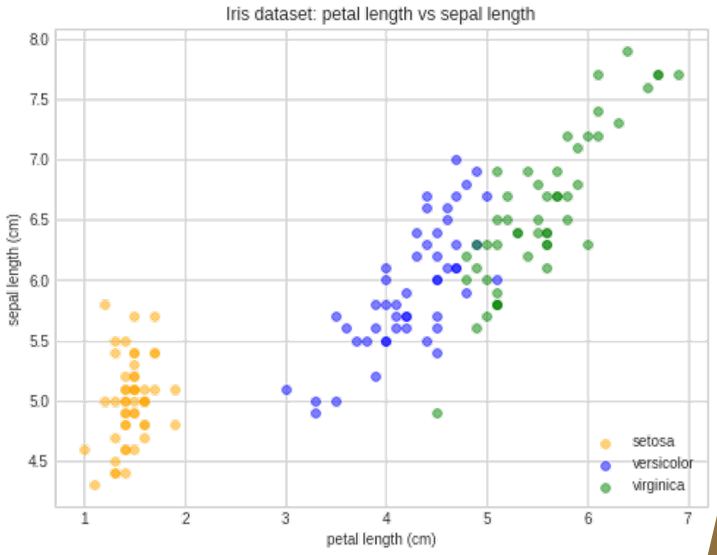
Bar plots

- Used for nominal or ordinal categories
- Compare data amongst different categories
- Ideal for more than 3 categories
- Can show large data changes over time



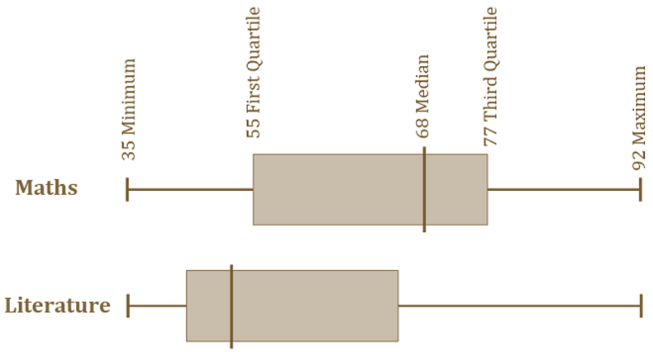
Scatter plots

- Used to visualize relation between two numeric variables
- Used to visualize correlation in a large data set
- Predict behavior of dependent variable based on the measure of the independent variable.



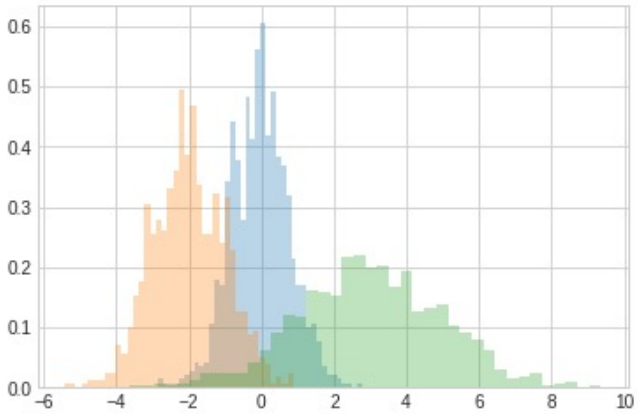
Box plots

- *aka* whisker plot
- Statistical graph used on sets of numerical data
- Shows the range, spread and center
- Used to compare data from different categories



Histograms

- Used for continuous data
- Displays the frequency distribution (shape)
- Summarize large data sets graphically
- Compare multiple distributions



LEARN BY DOING

To access the videos and material from the workshop series please visit:
https://guides.lib.purdue.edu/d-velop