Data Visualization: Python
MSL7320
IIT Jodhpur

Instructor: Revendra T

Outline

1. Pre-requisites
2. Motivations
3. Packages Ecosystem in R & Python
4. Data Visualization Tools
5. Applications of Data Visualization using Python
    - Univariate
    - Multivariate

Pre-requisites for Data Visualization

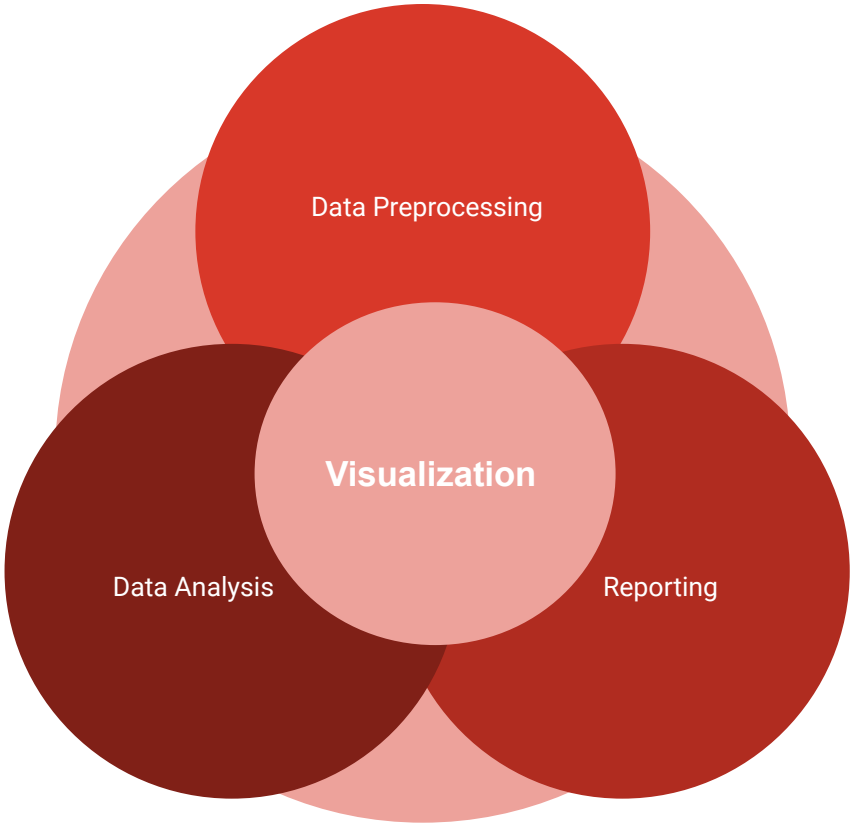
- Foundations in programming language (R/Python/Julia)
- Data pre-processing concepts
- Python:
  - Foundations of Python
  - Object oriented programming concepts
  - Numpy & Pandas libraries

Motivation for Data Visualization

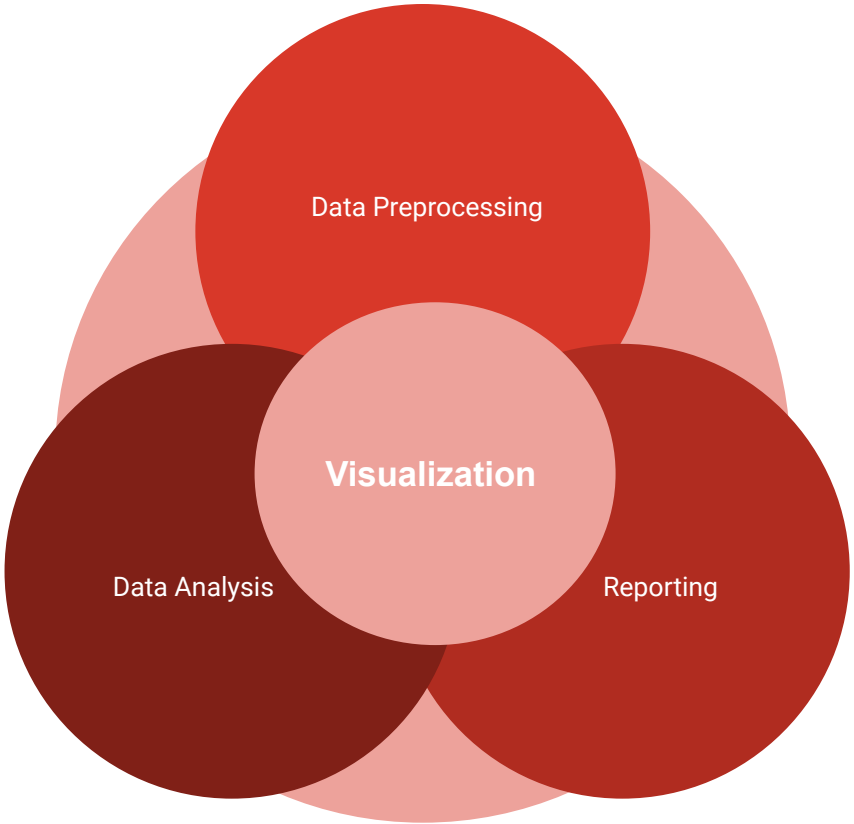
1. Data Preprocessing	2. Data Analysis	3. Reporting
• Data collection • Data organizing	• Statistical Modeling • Machine Learning	• **Visualization** • Publishing

Process in Business Analytics or Data Sciences

Motivation for Data Visualization

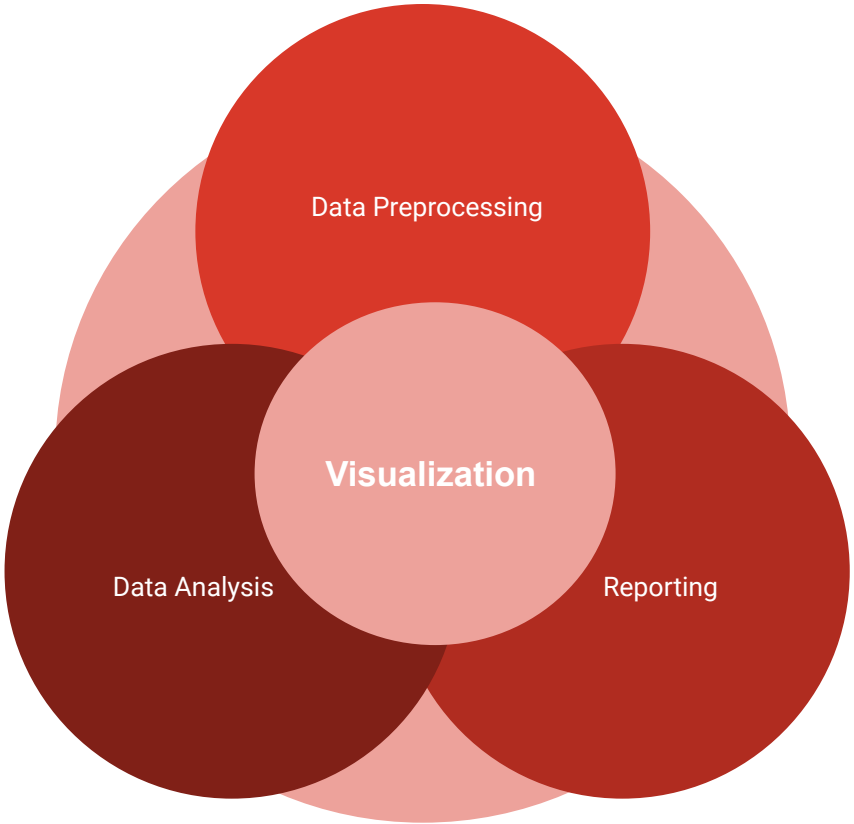


Motivation for Data Visualization



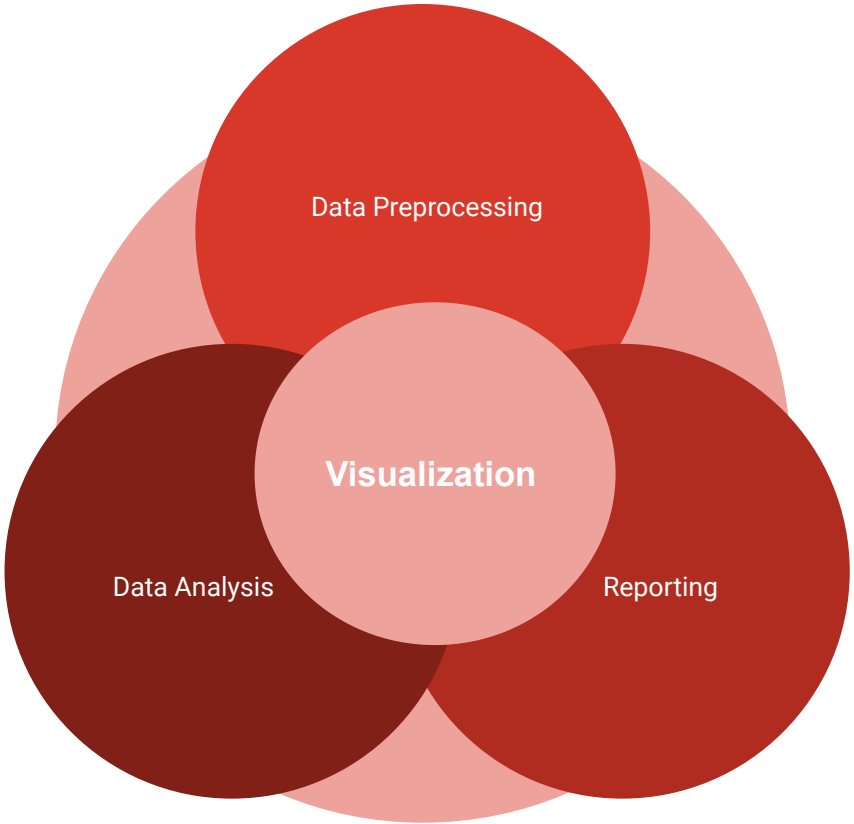
- Visualization is the heart of data sciences
- What does visualization solve in data sciences?

Motivation for Data Visualization



- Visualization is the heart of data sciences
- What does visualization solve in data sciences?

Motivation for Data Visualization

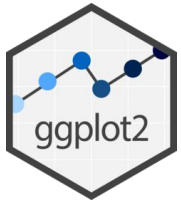


- Visualization is the heart of data sciences
- What does visualization solve in data sciences?
  - Part of communication to share findings or observations
  - Story-telling through visuals
  - Integrate in the process of decision-making
  - Communicate with target audience who are less familiar with statistics
  - Dashboards

Data Visualization: Programming

ggplot + extensions in R

- https://ggplot2.tidyverse.org/
- https://exts.ggplot2.tidyverse.org/gallery/



Data Visualization: Programming

ggplot + extensions in R

- https://ggplot2.tidyverse.org/
- https://exts.ggplot2.tidyverse.org/gallery/

Python

1. ggplot equivalent: plotnine
   a. https://plotnine.readthedocs.io/en/stable/
2. Matplotlib library
   a. https://matplotlib.org/
3. Seaborn library
   a. https://seaborn.pydata.org/
4. Pandas Plots
   a. https://pandas.pydata.org/docs/getting_started/intro_tutorials/04_plotting.html

Data Visualization: Tools

1. Tableau
   a. https://www.tableau.com/
2. PowerBI
   a. https://powerbi.microsoft.com/en-us/

More Tools:
https://www.g2.com/categories/data-visualization

Selecting Visualization Programming/Tool

1. Best match for business requirements
    a. Complex visualizations: use programming (Python/R/Julia)
    b. Take off the shelf: Tableau/PowerBI
2. Integration with data science products/solutions
3. Review the budget available
4. Learning curve
5. Skilled candidates in the market

Python Visualization Ecosystem

Library	Description
Matplotlib	● The foundational & the most advanced plotting library in Python ecosystem. ● Provides super flexibility to customize plots ● Requires a lot of lines of code to accomplish visualization
Seaborn	● Built on top of Matplotlib to simplify plotting ● Provides high-level interface to statistical plots and more choices of colors for visual appeal
Plotnine	● Follows grammar of graphics similar to ggplot ecosystem in R ● Useful for complex visualization as it uses layering approach
Pandas Plot	● Built on top of Matplotlib to simplify plotting and integrate with Pandas dataframes or series objects ● Easiest to use with a very few lines of code ● Useful for quick plots during data exploration along with data preparation

Data Types: Overview

Data Type	Description	Category
Integer	Whole numbers	Numeric
Float	Real numbers	
Complex number	Has both real and imaginary components	
String	Group of characters	Text
Boolean or logical	Logical True (1) or False (0)	Boolean or Logical
Nothing or Blank	No data scenario	None or Missing
Date	Capture the time, duration., etc	Date-Time
Categorical	Categorize information into levels.	Nominal, Ordinal

Data Type: In-built in Python; Not in-built in Python

Visualization: Components of a Plot

Components	Description
Title	• Title conveys the purpose of a visualization. • Answers a question . • Acts as a guide for the reader to understand the purpose of your visualization. • Suggests a possible conclusion.
Annotation	• Helps focus on some section of a plot. • Limit annotations to a few important sections of a plot
Legend	• Part of a plotting area • Explains the symbols used in the plot.
Objects	• Dots, lines. polygons etc to represent points and connections among points
Background	• Default or custom background

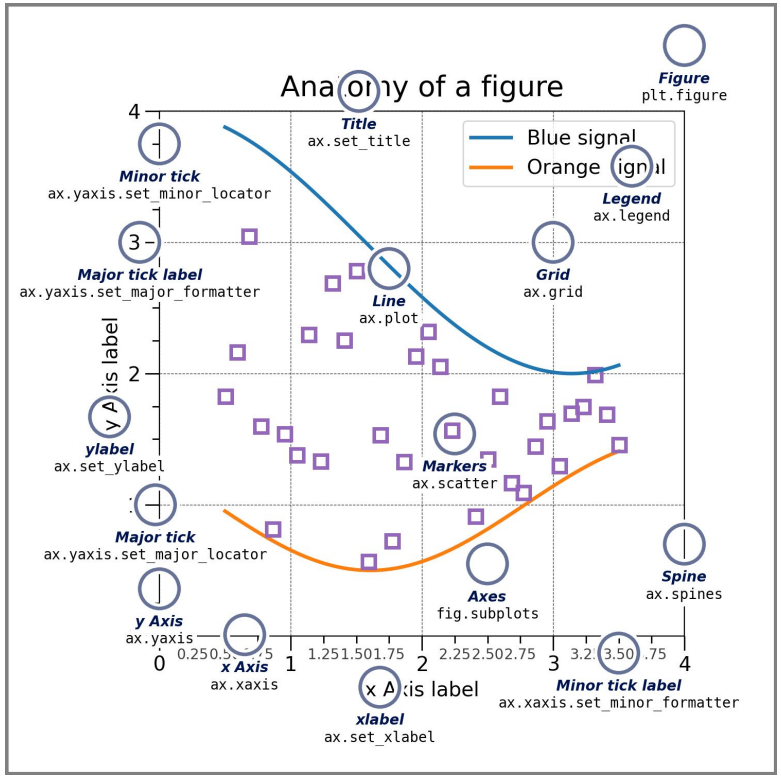
Reyes (2022)

Visualization: Components of a Plot

Components	Description
Color	• **Two to three variables**: Black & white, or gray scale • **Ordinal values**: Use hue with different levels of illumination • **Nominal values**: Use different color to differentiate each nominal value • **Brewer Palettes**: customize different colors
Margins	• Plot and the legend have margins, and they are within other margins.
Alignment	• The title, subtitle, source caption and axis titles alignment • **Horizontal**: left, right, & centre • **Vertical:** top & bottom
Reposition	• Re-position default to something in order such as ascending/descending
Erase	• Erase defaults such as lines, dots, axes etc and replace with custom options

Reyes (2022)

Illustration of Plot Components: Matplotlib Library



Practical Notebook: **Components of a plot.ipynb**

Visualization: Tabular Data

	Date	Ticker	Open	Close	Volume
0	2023-06-01	AAPL	250.10	252.30	1000000
1	2023-06-02	GOOGL	2100.50	2120.75	500000
2	2023-06-03	MSFT	300.25	305.40	750000

- What kind of insights do you extract from this tabular data?

Visualization: Tabular Data

	Date	Ticker	Open	Close	Volume
0	2023-06-01	AAPL	250.10	252.30	1000000
1	2023-06-02	GOOGL	2100.50	2120.75	500000
2	2023-06-03	MSFT	300.25	305.40	750000

Use plots in Pandas library
Illustration: **Visualize Single variable.ipynb**

- **What kind of insights do you extract from this tabular data?**
  - **Singular variable**
    - Raw values (close stock price)
    - Calculated values (daily returns)
    - Categorical values
  - **Multiple variables**
    - Categorical - categorical values
    - Numeric - categorical values
    - Numeric - numeric values

Visualization: Next Steps

- Non-tabular Data
  - Geospatial
  - Text
  - Network
  - Images
  - Video
  - Audio
- Mastery
  - Pandas Plots
  - Seaborn
  - Matplotlib
- Grammar of Graphics
  - ggplot extensions in R
  - plotnine in Python

Questions?

Connect:
revendra.iisc@gmail.com
Youtube.com/@revendrat

References

- Reyes, J. M. M. (2022). Data Visualization for Social and Policy Research: A Step-by-step Approach Using R and Python. Cambridge University Press.
- Read data from Yahoo Finance: https://pypi.org/project/yfinance/
- Convert multiple tickers to two-dimensional dataframe https://stackoverflow.com/questions/63107594/how-to-deal-with-multi-level-column-names-downloaded-with-yfinance/63107801#63107801
- Pandas plots: https://pandas.pydata.org/docs/getting_started/intro_tutorials/04_plotting.html