

DATA VISUALIZATION WITH PYTHON

Introduction to Data Science

Author: Eng. Carlos Andrés Sierra, M.Sc.
carlos.andres.sierra.v@gmail.com

Lecturer
Computer Engineer
School of Engineering
Universidad Distrital Francisco José de Caldas

2024-II



Outline

1 Principles of Information Visualization

2 Charting with Matplotlib

3 Visualizations with Seaborn



Outline

1 Principles of Information Visualization

2 Charting with Matplotlib

3 Visualizations with Seaborn

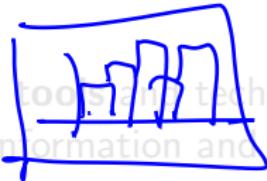


Introduction

- **Data visualization** is the **graphical representation** of information and data.
- By using **visual elements** like charts, graphs, and maps, data visualization tools provide an accessible way to see and **understand** trends, outliers, and patterns in data.
- In the world of Big Data, **data visualization** tools and technologies are essential to analyze massive amounts of information and make **data-driven decisions**.



rows

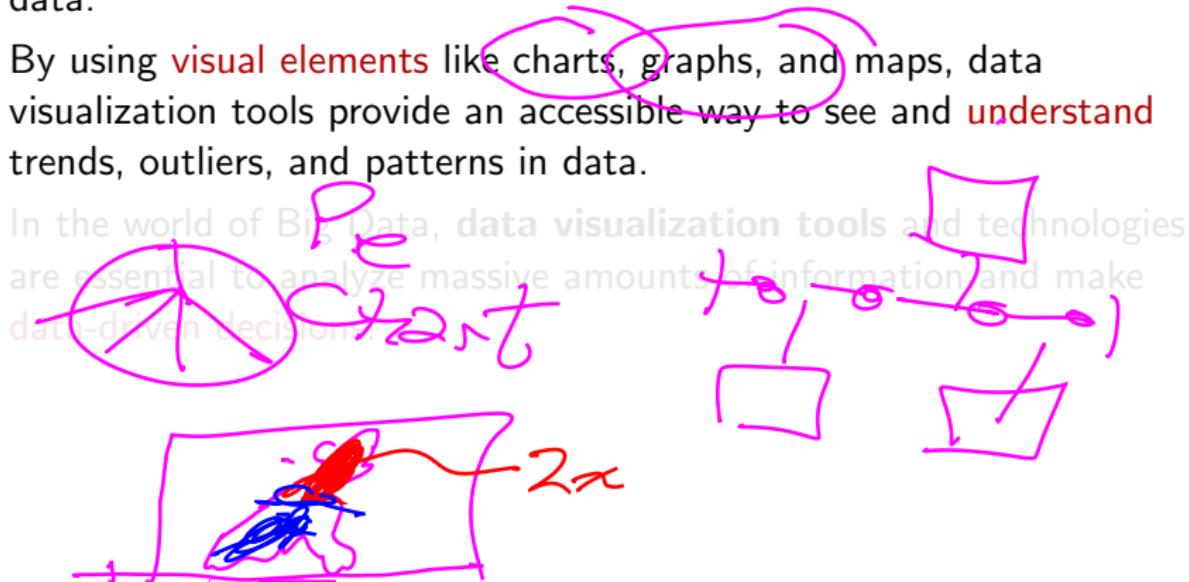


outliers



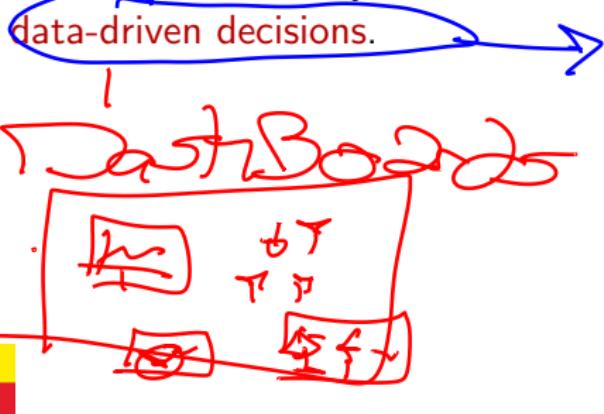
Introduction

- **Data visualization** is the **graphical representation** of information and data.
- By using **visual elements** like **charts, graphs, and maps**, data visualization tools provide an accessible way to see and **understand** trends, outliers, and patterns in data.
- In the world of Big Data, **data visualization tools** and technologies are essential to analyze massive amounts of information and make **data-driven decisions**.



Introduction

- **Data visualization** is the **graphical representation** of information and data.
- By using **visual elements** like charts, graphs, and maps, data visualization tools provide an accessible way to see and **understand** trends, outliers, and patterns in data.
- In the world of Big Data, **data visualization tools** and technologies are essential to analyze massive amounts of information and make **data-driven decisions**.

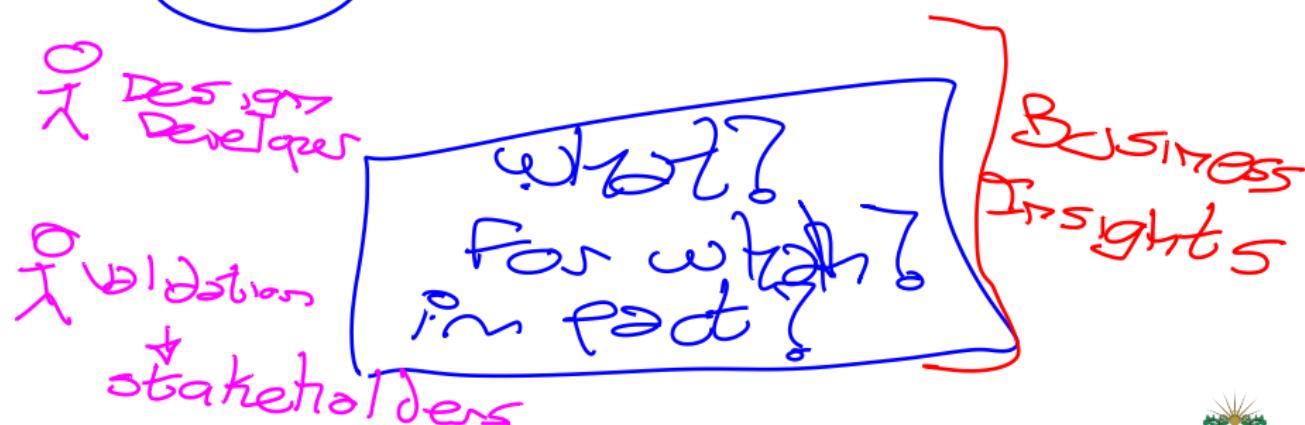


Santos
SCON PCS
New
Data
Data Decs.



Tools for Thinking About Design

- **Design** is the process of creating a plan or convention for the construction of an object or a system.
- **Design** has different connotations in different fields.
- In some cases, the direct construction of an object is also considered to be **design**.



Visualization Wheel

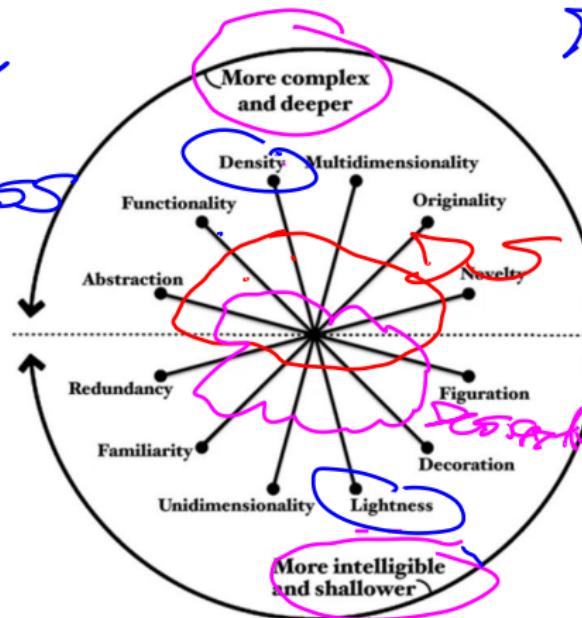
The **Visualization Wheel** is a tool that helps you to think about the design of your visualization.

Alberto
Cano

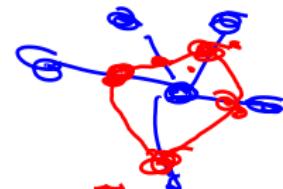
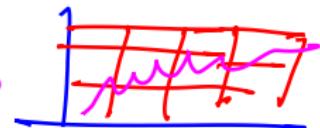
↑

~~THE TRUTH IS ELASTIC~~

Density
Lightness



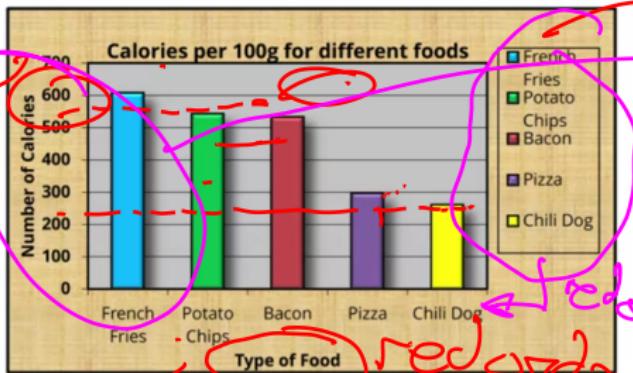
For consistency
of style
decoration



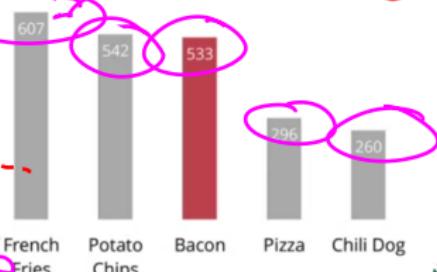
Graphical Heuristics

Edward Tufte

- Graphical heuristics are rules of thumb that help you to design effective visualizations.
- Graphical heuristics are based on the principles of perception and cognition.
- Graphical heuristics consist on pre-attentive processing, gestalt principles, and color theory.



Calories per 100g

~~Background~~

Comprehension and Memorability of Charts

- **Comprehension** is the ability to understand something.

- Memorability is the ability to remember something.

- Comprehension and memorability are important aspects of data visualization.

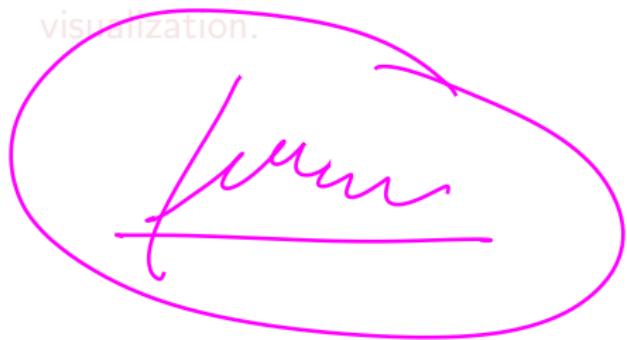


Comprehension



Comprehension and Memorability of Charts

- **Comprehension** is the ability to understand something.
- **Memorability** is the ability to remember something.
- Comprehension and memorability are important aspects of data visualization.



A handwritten note in pink ink. The word "Memorability" is written in cursive script, with a small five-pointed star drawn above the letter "a".



Comprehension and Memorability of Charts

- Comprehension is the ability to understand something.
- Memorability is the ability to remember something.
- Comprehension and memorability are important aspects of data visualization.

1. Time - likelihood

2. Avg. Confusion

3. Cost. → one interpretation



Colors in Data Visualization

RULES

INTUITIVENESS
Use intuitive colors. When choosing them, consider what associations do they evoke. If possible, use colors that audience will associate with your data anyway.

MODERATION
Use colors in moderation. For a simple dataset, a single color is preferable. Use color as a strategic tool to highlight the important parts of your visual.

CONSISTENCY
Use colors consistently. Change colors if you want your audience to feel the change for the specific reason, but never simply for the sake of novelty.

CLARITY
Use colors to make the data easier to read. Make sure your audience will be able to distinguish between the items shown in the visualization.

CLASSIFICATION
Don't use a gradient color palette for categories. And the other way round - different colors for same measurement.

EXPLAINABILITY
Make sure to explain to your audience what exactly used colors mean. Remember to create a color key.



Color Schemas

Monochromatic - the simplest formula for harmony is monochromatic. Consists of different shades of one hue. Not a great choice if we want to highlight something.



Analogous - this scheme is composed of colors that are next to each other on the wheel. Usually they match up pretty well, making elegant and clear look.



Complementary - uses two colors which are opposite on the color wheel. With saturated colors makes very vibrant look. Try to tone down colors to avoid overvibrance, by adjusting saturation and lightness/darkness. Do not use with text with saturated colours.



Triadic - uses three colors that evenly spaced on the color wheel. Makes that none of colors is dominant and quite vibrant look.



Split-Complementary - variation of complementary scheme. Uses base color and two adjacent to its complementary color. Often this scheme is more pleasant to the eye than usual complementary scheme.



Tetradic - this scheme consists of four colors, two of them are complementary to other two. Choosing one color as dominant and the rest as accents, gives the best result.



Remember! Do not stick strictly to colors imposed by a scheme. These patterns are just starting points, you can create your own variations based on schemes above. Check also: paletton.com



Color Palletes

QUANTITATIVE DATA - SEQUENTIAL OR DIVERGING COLORS

Color is used to show variations in the data. The palette contains a sequence of colors that clearly indicate which values are larger or smaller than which other ones (sequential scale). It can also visualize the deviation of data values in one of two directions relative to a neutral midpoint (diverging scale). Diverging scale can be viewed as two merged sequential scales.

| | Sequential scales | Diverging scales |
|---------|-------------------|------------------|
| Blues | | |
| Viridis | | |
| YlOrBr | | |

CATEGORICAL DATA - QUALITATIVE COLORS

Color is used to separate areas into distinct categories. The palette should consist of colors as distinct from one another as possible. The maximum number of categories that can be displayed is about 12 (practically speaking, probably fewer).

| | Colorblind | Accent | Pastell |
|-------|------------|--------|---------|
| Set12 | | | |

All examples are available in Seaborn library. Check also: matplotlib.github.io/_wianthue/

USAGE GUIDELINES

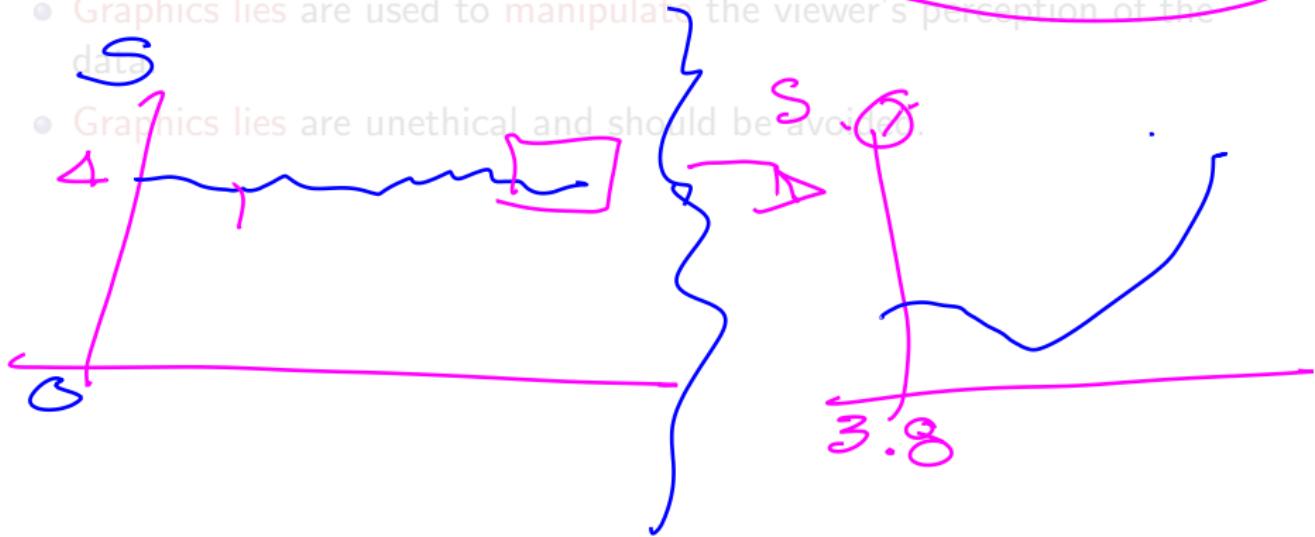
| SCALES | Sequential | Diverging | Categorical |
|-------------|--|---|--|
| Categorical | <input checked="" type="checkbox"/> A B C D E | <input checked="" type="checkbox"/> A B C D E | <input checked="" type="checkbox"/> A B C D E |
| Ordinal | <input checked="" type="checkbox"/> Low High | <input checked="" type="checkbox"/> Low Medium High | <input checked="" type="checkbox"/> Low Medium High |
| Interval | <input checked="" type="checkbox"/> 1 2 3 4 5 6 | <input checked="" type="checkbox"/> 1 2 3 4 5 6 | <input checked="" type="checkbox"/> 1-2 3-4 5-6 |
| Ratio | <input checked="" type="checkbox"/> -5 -3 -1 1 3 5 | <input checked="" type="checkbox"/> 5 -3 -1 1 3 5 | <input checked="" type="checkbox"/> 6 to -2 -2 to 2 2 to 6 |

best
red
blue
green
yellow
purple
orange
pink
grey
black
white



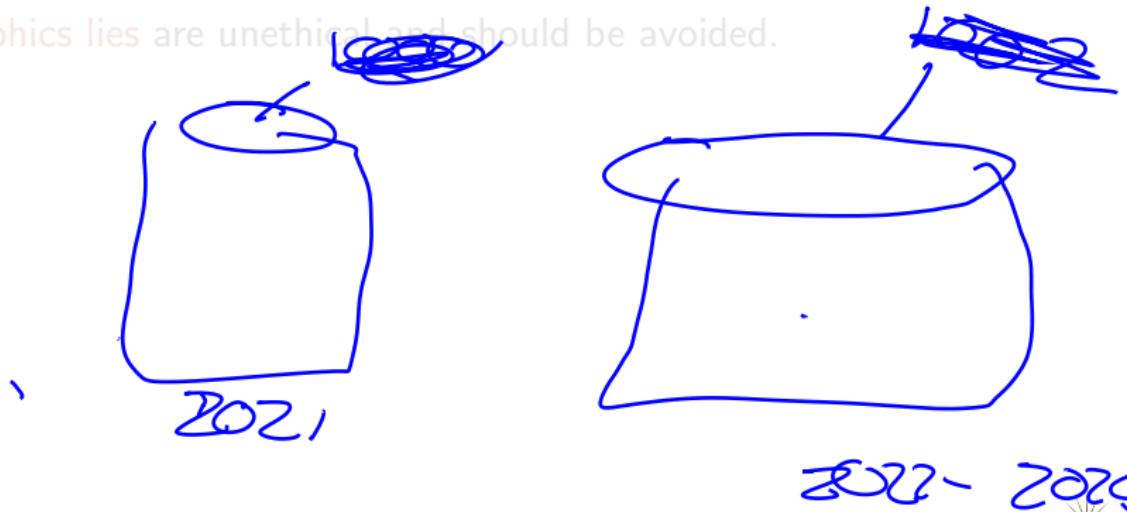
Graphics Lies, Misleading Visuals

- Graphics lies are visualizations that are misleading or deceptive.
- Graphics lies are used to manipulate the viewer's perception of the data.
- Graphics lies are unethical and should be avoided.



Graphics Lies, Misleading Visuals

- Graphics lies are visualizations that are **misleading** or **deceptive**.
- Graphics lies are used to **manipulate** the viewer's perception of the data.
- Graphics lies are unethical and should be avoided.



Graphics Lies, Misleading Visuals

- Graphics lies are visualizations that are **misleading** or **deceptive**.
- Graphics lies are used to **manipulate** the viewer's perception of the data.
- **Graphics lies** are unethical and should be avoided.



Outline

1 Principles of Information Visualization

2 Charting with Matplotlib

3 Visualizations with Seaborn



What is Matplotlib?

- **Matplotlib** is a comprehensive **library** for creating static, animated, and interactive **visualizations in Python**.
- **Matplotlib** makes easy things easy and hard things **possible**
- **Matplotlib** can be used in Python scripts, the Python and IPython shell, web application servers, and various graphical user interface toolkits.

Matlab \Rightarrow Canvas
objects



Matplotlib Architecture

- Matplotlib is a multi-platform data visualization library built on NumPy arrays.
- Matplotlib is designed to work with the broader SciPy stack.
- Matplotlib is a 2D plotting library that produces publication-quality figures in a variety of formats.
- Matplotlib has some layers that are responsible for different aspects of the visualization process.



free
SciPy

API

→ Pandas → serve Dataframe → NumPy Array



Matplotlib Architecture

- **Matplotlib** is a multi-platform data visualization library built on NumPy arrays.
- **Matplotlib** is designed to work with the broader SciPy stack.
- Matplotlib is a 2D plotting library that produces publication-quality figures in a variety of formats.
- Matplotlib has some layers that are responsible for different aspects of the visualization process.



Matplotlib Architecture

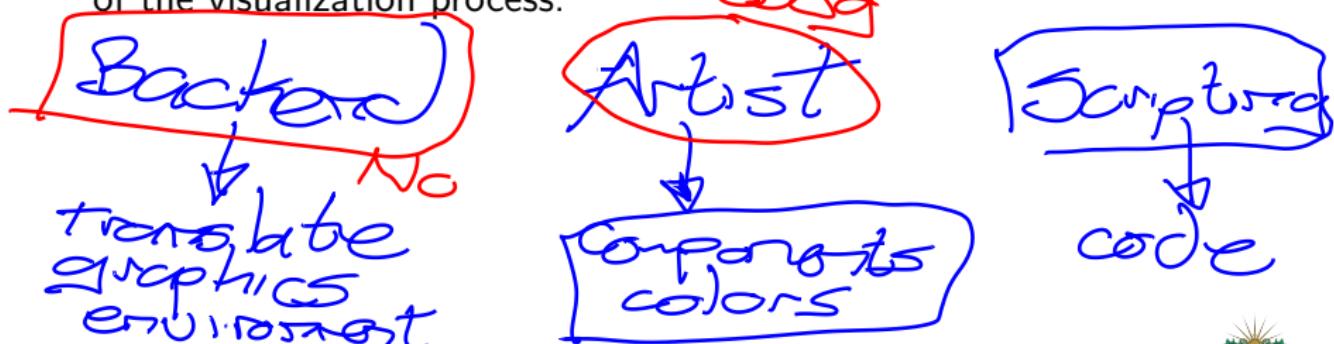
- **Matplotlib** is a multi-platform data visualization library built on NumPy arrays.
- **Matplotlib** is designed to work with the broader SciPy stack.
- **Matplotlib** is a 2D plotting library that produces publication-quality figures in a variety of formats.
- Matplotlib has some layers that are responsible for different aspects of the visualization process.

PNG → 720 × 3024



Matplotlib Architecture

- **Matplotlib** is a multi-platform data visualization library built on NumPy arrays.
- **Matplotlib** is designed to work with the broader SciPy stack.
- **Matplotlib** is a 2D plotting library that produces publication-quality figures in a variety of formats.
- **Matplotlib** has some layers that are responsible for different aspects of the visualization process.



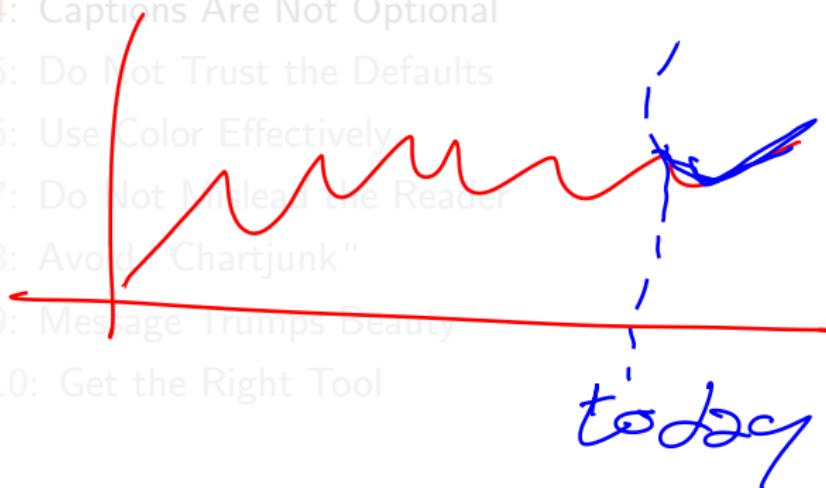
Ten Simple Rules for Better Figures

- Rule 1: Know Your Audience
- Rule 2: Identify Your Message
- Rule 3: Adapt the Figure to the Support Medium
- Rule 4: Captions Are Not Optional
- Rule 5: Do Not Trust the Defaults
- Rule 6: Use Color Effectively
- Rule 7: Do Not Mislead the Reader
- Rule 8: Avoid "Chartjunk"
- Rule 9: Message Trumps Beauty
- Rule 10: Get the Right Tool



Ten Simple Rules for Better Figures

- Rule 1: Know Your Audience
- Rule 2: Identify Your Message
- Rule 3: Adapt the Figure to the Support Medium
- Rule 4: Captions Are Not Optional
- Rule 5: Do Not Trust the Defaults
- Rule 6: Use Color Effectively
- Rule 7: Do Not Mislead the Reader
- Rule 8: Avoid "Chartjunk"
- Rule 9: Message Trumps Beauty
- Rule 10: Get the Right Tool



Ten Simple Rules for Better Figures

- Rule 1: Know Your Audience
 - Rule 2: Identify Your Message
 - Rule 3: Adapt the Figure to the Support Medium
 - Rule 4: Captions Are Not Optional
 - Rule 5: Do Not Trust the Defaults
 - Rule 6: Use Color Effectively
 - Rule 7: Do Not Mislead the Reader
 - Rule 8: Avoid "Chartjunk"
 - Rule 9: Message Trumps Beauty
 - Rule 10: Get the Right Tool
- WG => code screen*
GIF => P.X.

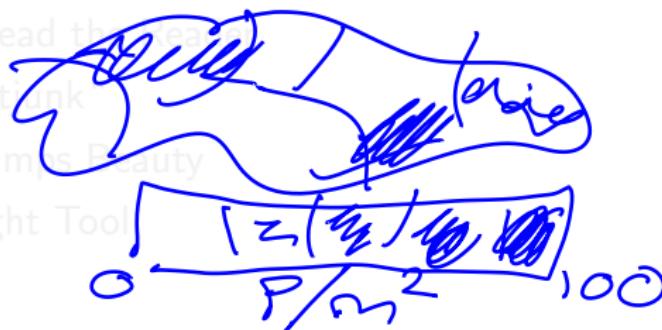


Ten Simple Rules for Better Figures

- Rule 1: Know Your Audience
- Rule 2: Identify Your Message
- Rule 3: Adapt the Figure to the Support Medium
- Rule 4: Captions Are Not Optional

- Rule 5: Do Not Trust the Defaults
- Rule 6: Use Color Effectively
- Rule 7: Do Not Mislead the Reader
- Rule 8: Avoid "Chartjunk"
- Rule 9: Message Trumps Beauty
- Rule 10: Get the Right Tool

Population



Ten Simple Rules for Better Figures

- Rule 1: Know Your Audience
- Rule 2: Identify Your Message
- Rule 3: Adapt the Figure to the Support Medium
- Rule 4: Captions Are Not Optional
- Rule 5: Do Not Trust the Defaults
- Rule 6: Use Color Effectively
- Rule 7: Do Not Mislead the Reader
- Rule 8: Avoid "Chartjunk"
- Rule 9: Message Trumps Beauty
- Rule 10: Get the Right Tool

Defaults
Parameters



Ten Simple Rules for Better Figures

- Rule 1: Know Your Audience
- Rule 2: Identify Your Message
- Rule 3: Adapt the Figure to the Support Medium
- Rule 4: Captions Are Not Optional
- Rule 5: Do Not Trust the Defaults
- Rule 6: Use Color Effectively
- Rule 7: Do Not Mislead the Reader
- Rule 8: Avoid "Chartjunk"
- Rule 9: Message Trumps Beauty
- Rule 10: Get the Right Tool

color theory
graphics heuristics



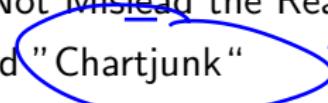
Ten Simple Rules for Better Figures

- Rule 1: Know Your Audience
- Rule 2: Identify Your Message
- Rule 3: Adapt the Figure to the Support Medium
- Rule 4: Captions Are Not Optional
- Rule 5: Do Not Trust the Defaults
- Rule 6: Use Color Effectively
- Rule 7: Do Not Mislead the Reader
- Rule 8: Avoid "Chartjunk"
- Rule 9: Message Trumps Beauty
- Rule 10: Get the Right Tool

No
Lies



Ten Simple Rules for Better Figures

- Rule 1: Know Your Audience
- Rule 2: Identify Your Message
- Rule 3: Adapt the Figure to the Support Medium
- Rule 4: Captions Are Not Optional
- Rule 5: Do Not Trust the Defaults
- Rule 6: Use Color Effectively
- Rule 7: Do Not Mislead the Reader
- Rule 8: Avoid "Chartjunk"
- Rule 9: Message Trumps Beauty
- Rule 10: Get the Right Tool

add grids
add background good



Ten Simple Rules for Better Figures

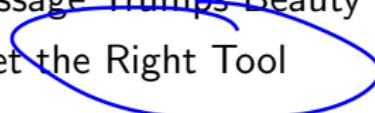
- Rule 1: Know Your Audience
- Rule 2: Identify Your Message
- Rule 3: Adapt the Figure to the Support Medium
- Rule 4: Captions Are Not Optional
- Rule 5: Do Not Trust the Defaults
- Rule 6: Use Color Effectively
- Rule 7: Do Not Mislead the Reader
- Rule 8: Avoid "Chartjunk"
- Rule 9: Message Trumps Beauty
- Rule 10: Get the Right Tool

No saturate
Simple / clear



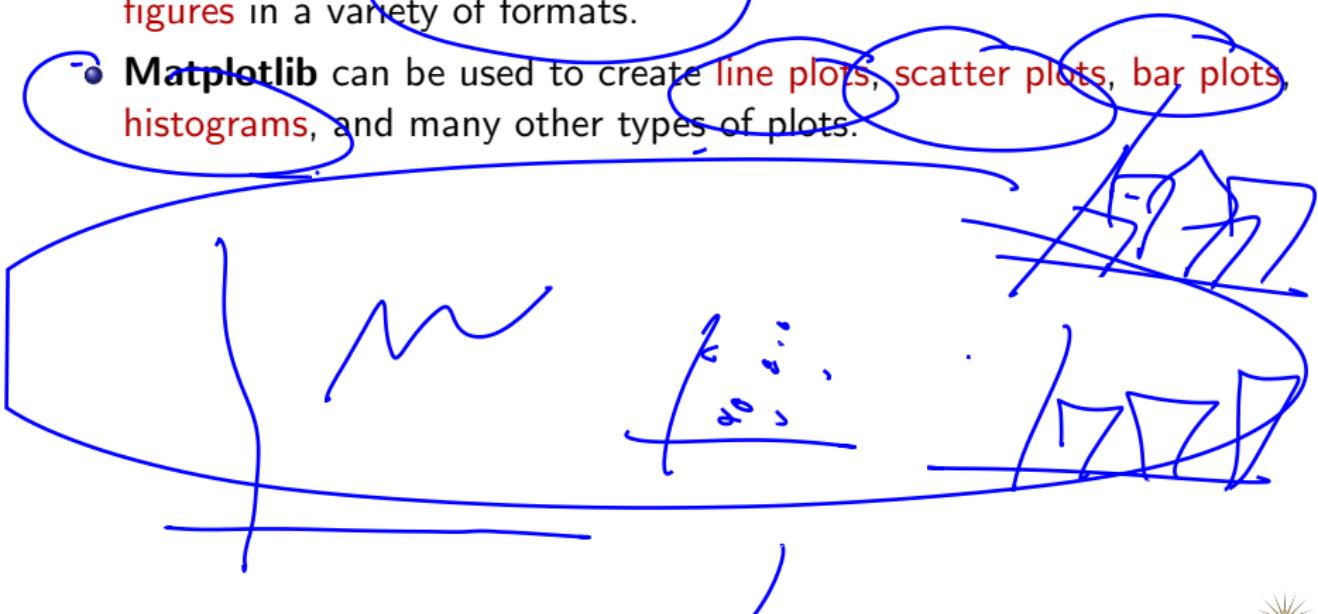
Ten Simple Rules for Better Figures

- Rule 1: Know Your Audience
- Rule 2: Identify Your Message
- Rule 3: Adapt the Figure to the Support Medium
- Rule 4: Captions Are Not Optional
- Rule 5: Do Not Trust the Defaults
- Rule 6: Use Color Effectively
- Rule 7: Do Not Mislead the Reader
- Rule 8: Avoid "Chartjunk"
- Rule 9: Message Trumps Beauty
- Rule 10: Get the Right Tool



Basic Plotting with Matplotlib

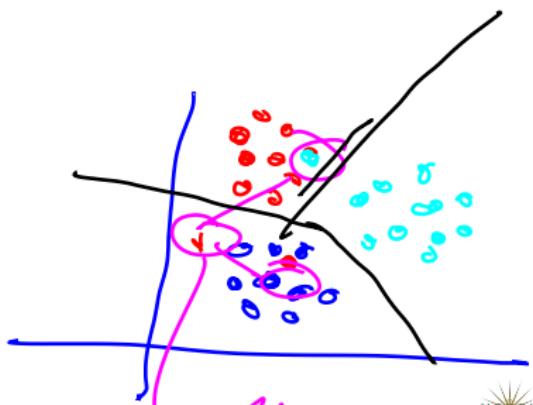
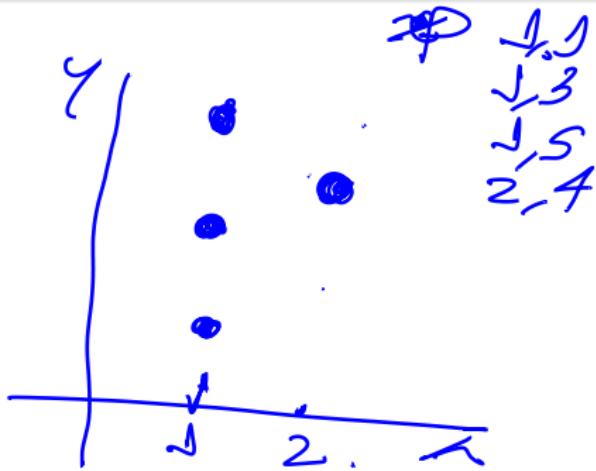
- **Matplotlib** is a **2D plotting library** that produces **publication-quality figures** in a variety of formats.
- **Matplotlib** can be used to create **line plots**, **scatter plots**, **bar plots**, **histograms**, and many other types of plots.



Scatter Plots

Definition

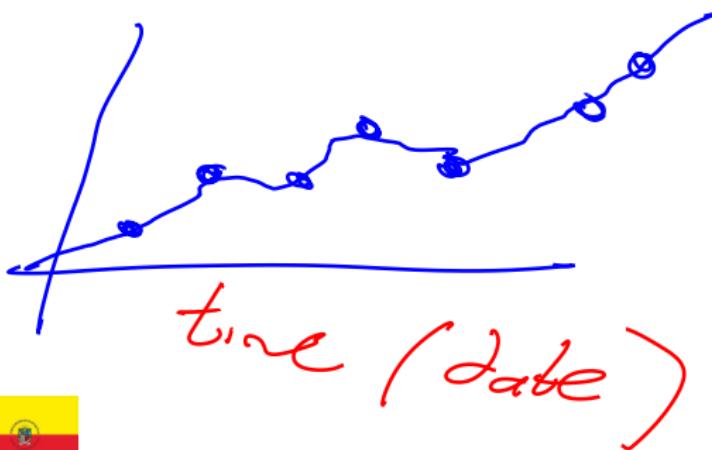
A **scatter plot** is a type of plot that shows individual data points along the **x** and **y** axis. Scatter plots are used to observe **relationships** between variables.



Line Plots

Definition

A **line plot** is a type of plot that displays information as a series of data points called **markers** connected by straight line segments. **Line plots** are used to observe the **trend** in data over intervals of time.



Bar Plots

Definition

A **bar plot** is a type of plot that presents **categorical data** with rectangular bars with lengths proportional to the values that they represent. **Bar plots** are used to compare **quantitative** data across different categories.



Subplots

- **Subplots** are groups of smaller axes that can exist together within a single figure.
- **Subplots** are useful when you want to show multiple plots in the same figure.
- **Subplots** are created using the `plt.subplots()` function.



Histograms

Definition

A **histogram** is a type of plot that displays the **distribution** of a dataset. **Histograms** are used to show the **frequency** of values in a dataset.



Box Plots

Definition

A **box plot** is a type of plot that displays the **distribution** of a dataset. **Box plots** are used to show the **spread** and **central tendency** of values in a dataset.



HeatMaps

Definition

A **heatmap** is a type of plot that displays the **intensity** of data at the intersection of two variables. **Heatmaps** are used to show the **correlation** between variables in a dataset.



Animation

- **Animation** is the process of creating a sequence of images that change over time.
- **Animation** is created using the `FuncAnimation` class from the `matplotlib.animation` module.



Widgets

- Widgets can be used to create **interactive visualizations** in a **Jupyter notebook**.
- Widgets can be used to create **sliders**, **buttons**, and other interactive elements.
- Widgets are created using the `ipywidgets` library.



Widgets

- Widgets can be used to create **interactive visualizations** in a **Jupyter notebook**.
- Widgets can be used to create **sliders**, **buttons**, and other interactive elements.
- Widgets are created using the `ipywidgets` library.



Widgets

- Widgets can be used to create **interactive visualizations** in a **Jupyter notebook**.
- Widgets can be used to create **sliders**, **buttons**, and other interactive elements.
- Widgets are created using the `ipywidgets` library.



Outline

1 Principles of Information Visualization

2 Charting with Matplotlib

3 Visualizations with Seaborn



What is Seaborn?

- **Seaborn** is a Python data visualization library based on **Matplotlib**.
- **Seaborn** provides a high-level interface for creating **attractive** and **informative** statistical graphics.
- **Seaborn** is built on top of **Matplotlib** and closely integrated with the **data structures** from **pandas**.



Spurious Correlations

Definition

A **spurious correlation** is a **statistical relationship** between two variables that is **not causally related**. Spurious correlations are often the result of **confounding variables**.



Mapping and Geographic

Definition

Mapping is the process of creating a **visual representation** of an area.

Mapping is used to show the **geographic** distribution of data.



Outline

1 Principles of Information Visualization

2 Charting with Matplotlib

3 Visualizations with Seaborn



Thanks!

Questions?



Repo: <https://github.com/EngAndres/ud-public/tree/main/courses/data-science-introduction>

