

Practical Business Python

Lecture 7: Data Visualization with Python.

Igor Vyshnevskiy

Woosong University

October 26, 2023

Agenda

1. Introduction to Data Visualization
2. Importance of Data Visualization
3. Chart Types
4. Visual Cues
5. Libraries
6. Examples
7. Visualization Practicum
8. In-class assignment

1. Introduction to Data Visualization

What is Data Visualization

- ***Data visualization*** is the process of representing data and information in a graphical format.
- The goal of data visualization is to communicate insights and patterns in a more effective and meaningful way.
- Data visualization allows analysts, researchers, and decision-makers to easily understand complex data sets.
- Effective data visualization leverages design principles such as color, shape, and layout to make information more accessible and understandable.
- Data visualization enables users to quickly and efficiently gain insights from data.

2. Importance of Data Visualization

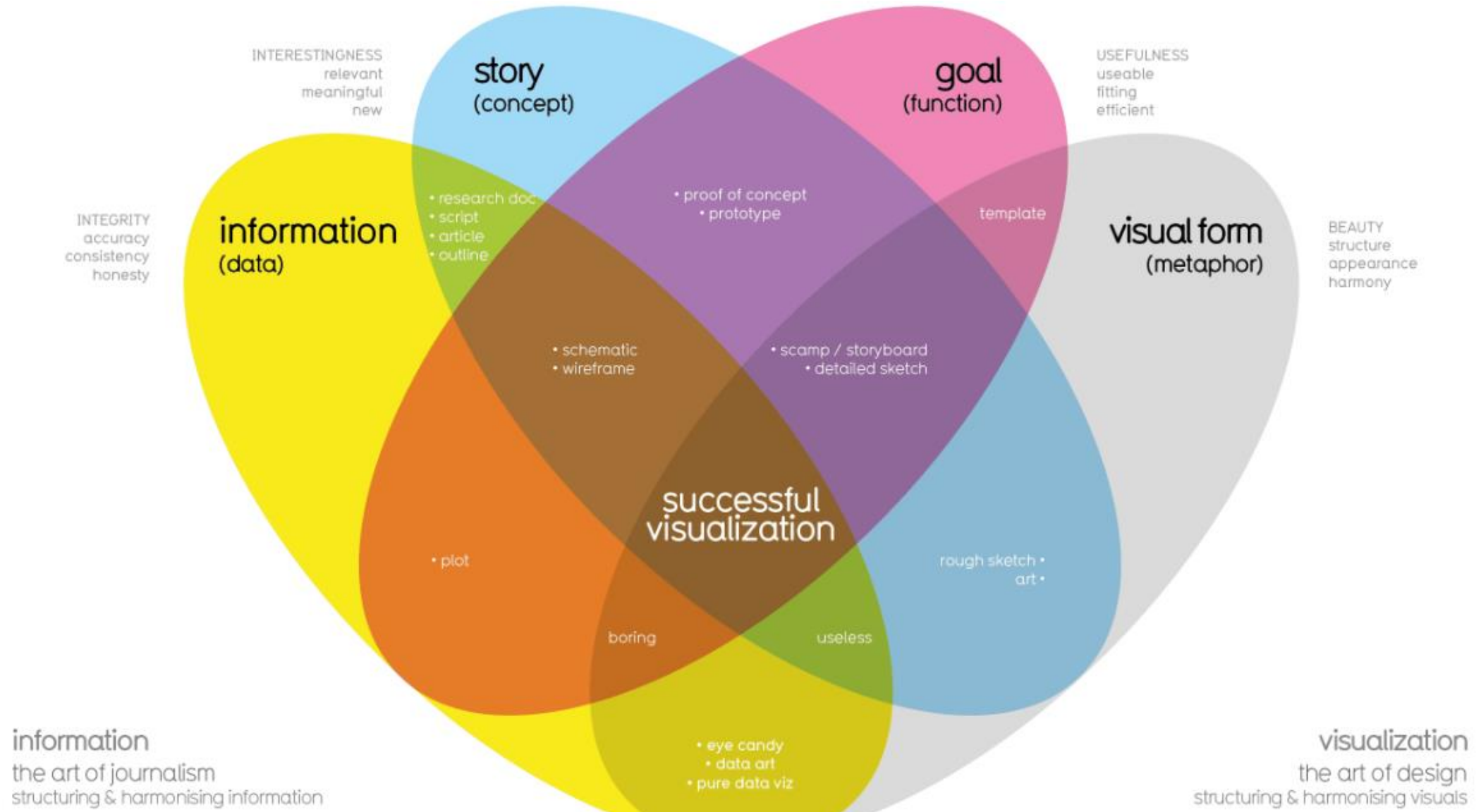
Effective data visualization:

- helps users to better understand patterns, trends, and relationships in data
- helps to identify outliers and anomalies in data that might be missed otherwise
- reveal hidden insights and relationships that are not immediately apparent in raw data
- helps to communicate findings and insights to stakeholders and decision-makers in a clear and compelling way
- helps to support data-driven decision-making by providing an intuitive and accessible view of data.

Poor data visualization can lead to:

- misleading interpretations and conclusions;
- oversimplification or obscuring of important details;
- confusion or misinterpretation of the data;
- biases or misrepresentations based on design choices;
- difficulty in visualizing certain types of data effectively;
- inaccessibility for users with visual impairments;
- incomplete or insufficient analysis due to a lack of context or nuance.

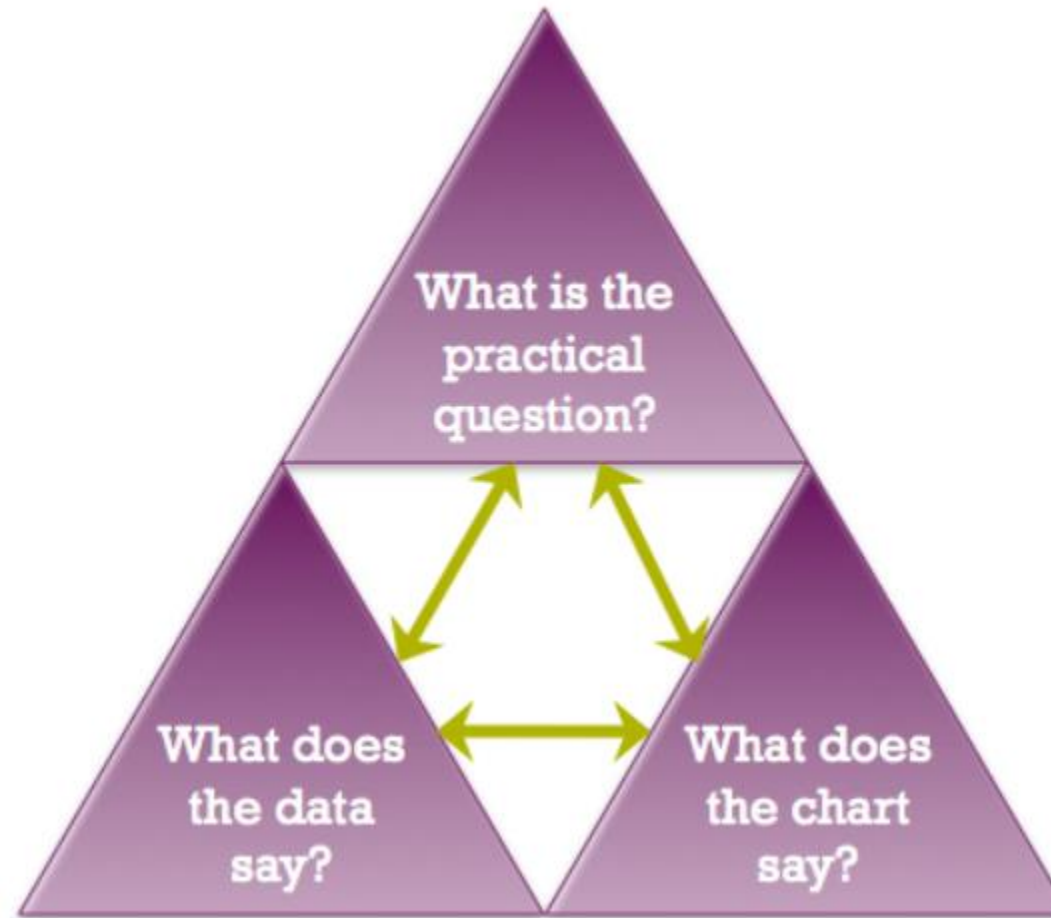
Four Elements of Good Data Visualization under the David McCandless method



- **Information (data):** The information or data that you are trying to convey is a key building block for your data visualization. Without information or data, you cannot communicate your findings successfully.
- **Story (concept):** Story allows you to share your data in meaningful and interesting ways. Without a story, your visualization is informative, but not really inspiring.
- **Goal (function):** The goal of your data visualization makes the data useful and usable. This is what you are trying to achieve with your visualization. Without a goal, your visualization might still be informative, but can't generate actionable insights.
- **Visual form (metaphor):** The visual form element is what gives your data visualization structure and makes it beautiful. Without visual form, your data is not visualized yet.

Kaiser Fung's Junk Charts Trifecta Checkup

to estimate the effectiveness of data visualization



What to Avoid

Cutting off the y-axis	Changing the scale on the y-axis can make the differences between different groups in your data seem more dramatic, even if the difference is actually quite small.
Misleading use of a dual y-axis	Using a dual y-axis without clearly labeling it in your data visualization can create extremely misleading charts.
Artificially limiting the scope of the data	If you only consider the part of the data that confirms your analysis, your visualizations will be misleading because they don't take all of the data into account.
Problematic choices in how data is binned or grouped	It is important to make sure that the way you are grouping data isn't misleading or misrepresenting your data and disguising important trends and insights.
Using part-to-whole visuals when the totals do not sum up appropriately	If you are using a part-to-whole visual like a pie chart to explain your data, the individual parts should add up to equal 100%. If they don't, your data visualization will be misleading.
Hiding trends in cumulative charts	Creating a cumulative chart can disguise more insightful trends by making the scale of the visualization too large to track any changes over time.
Artificially smoothing trends	Adding smooth trend lines between points in a scatterplot can make it easier to read that plot, but replacing the points with just the line can actually make it appear that the point is more connected over time than it actually was.

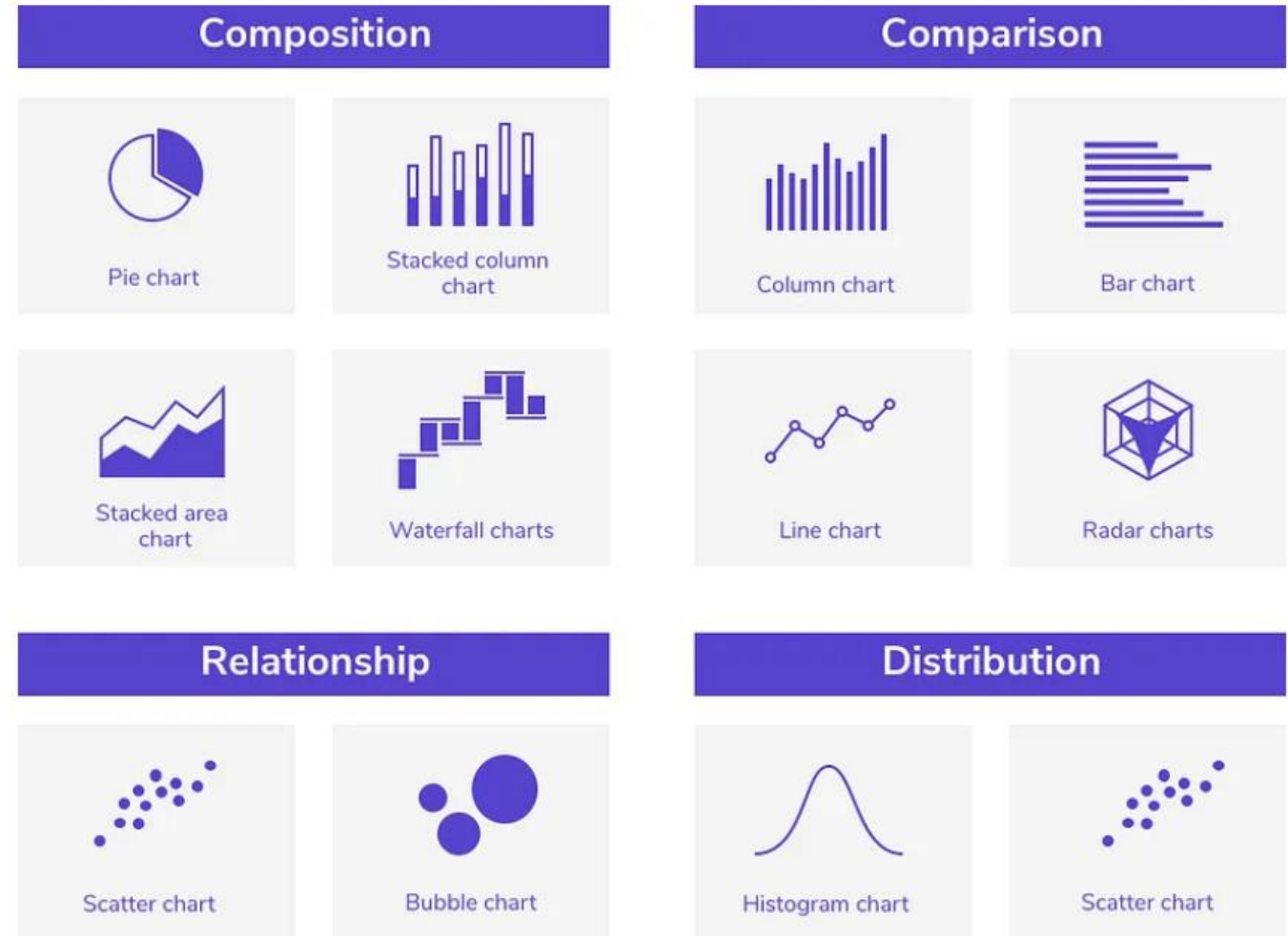
3. Chart Types

- Different types of visualizations are better suited to different types of data and communication goals
- Choosing the right visualization can help you communicate your insights more effectively and support decision-making.

The example of detail interactive decision tree to make decisions based on key questions that you can ask yourself I highly recommend:

<https://www.data-to-viz.com/>

The Most Common Chart Types



Source: <https://uxplanet.org/data-heavy-applications-how-to-design-perfect-charts-c0c893fef6de>

4. Visual Cues

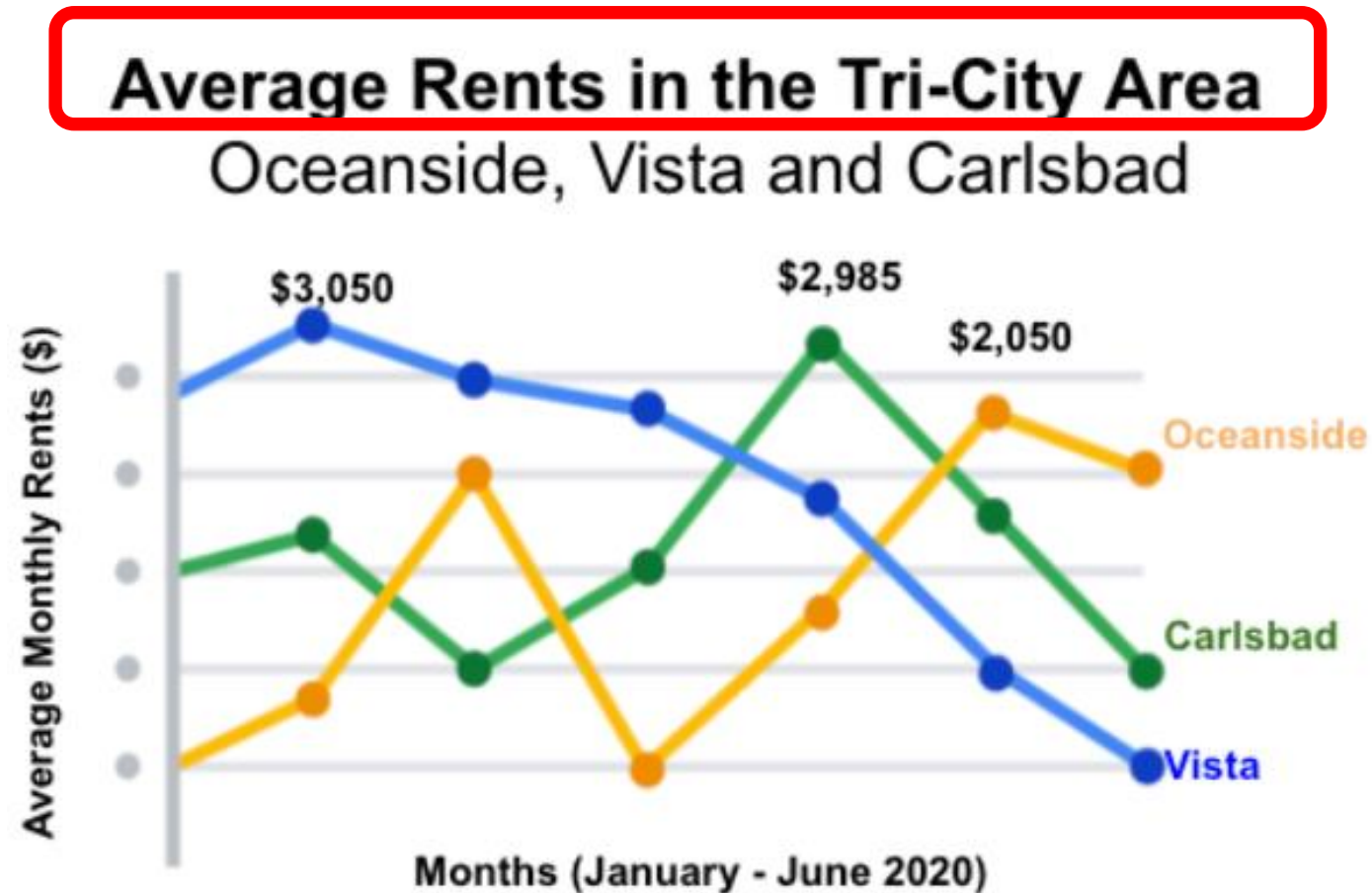
If you want to invite your audience into your presentation and keep them engaged, you have only **5 seconds** to catch their interest.

They should be able to process and understand the information you are trying to share with this extremely short time frame.

Effective visual cues are highly valuable for this purpose.

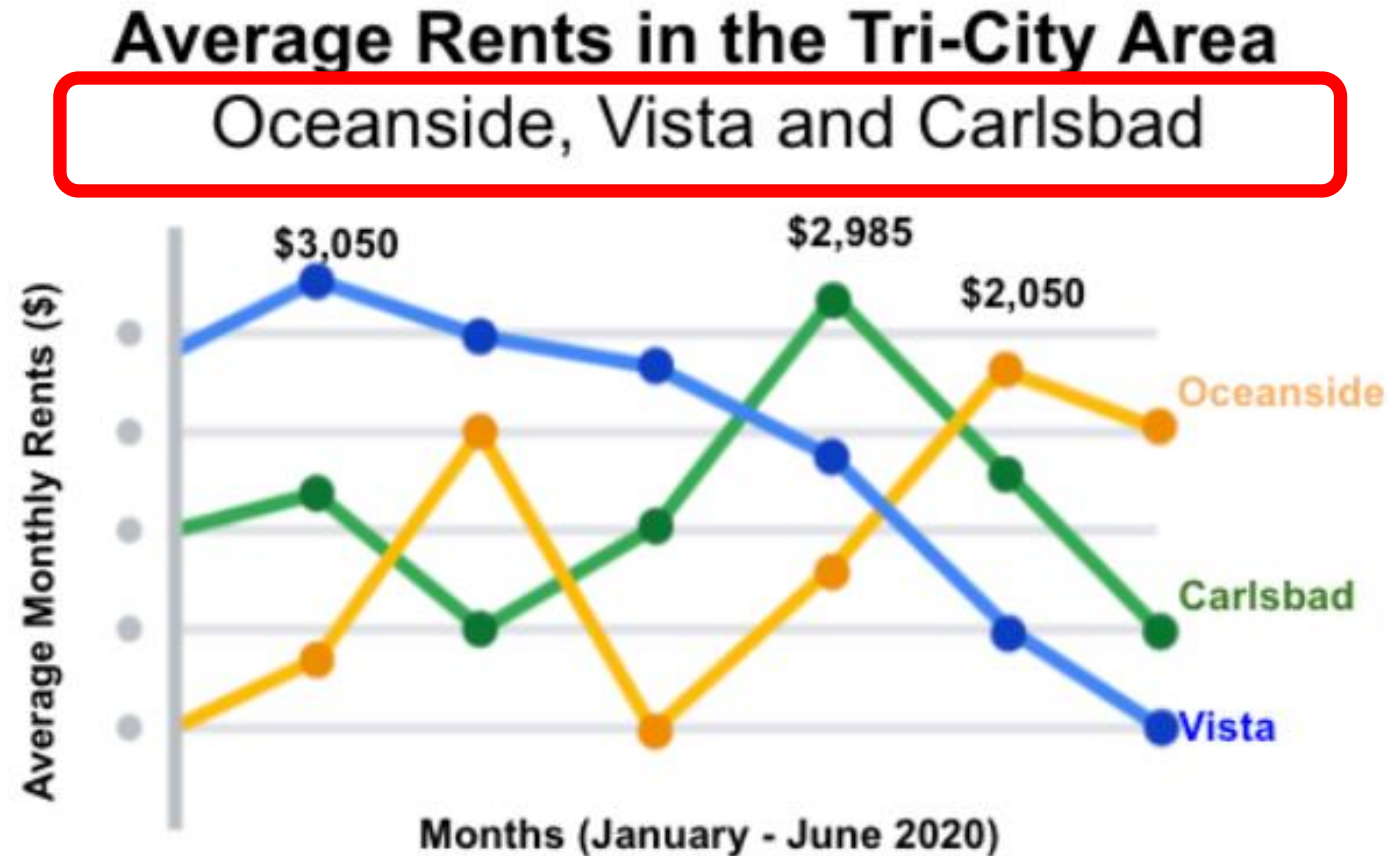
Effective Visual Cues

- Headlines that pop



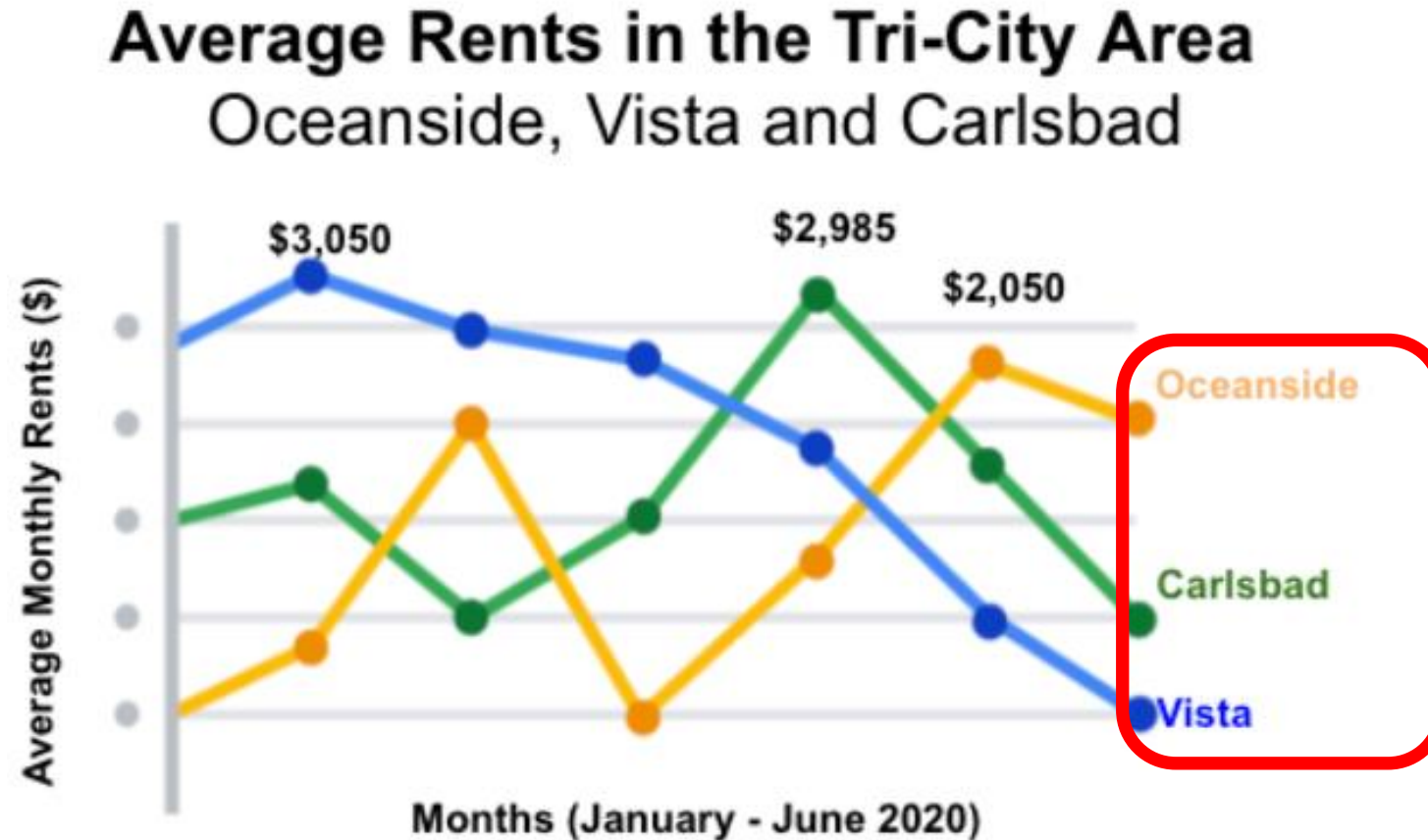
Effective Visual Cues

- Subtitles that clarify



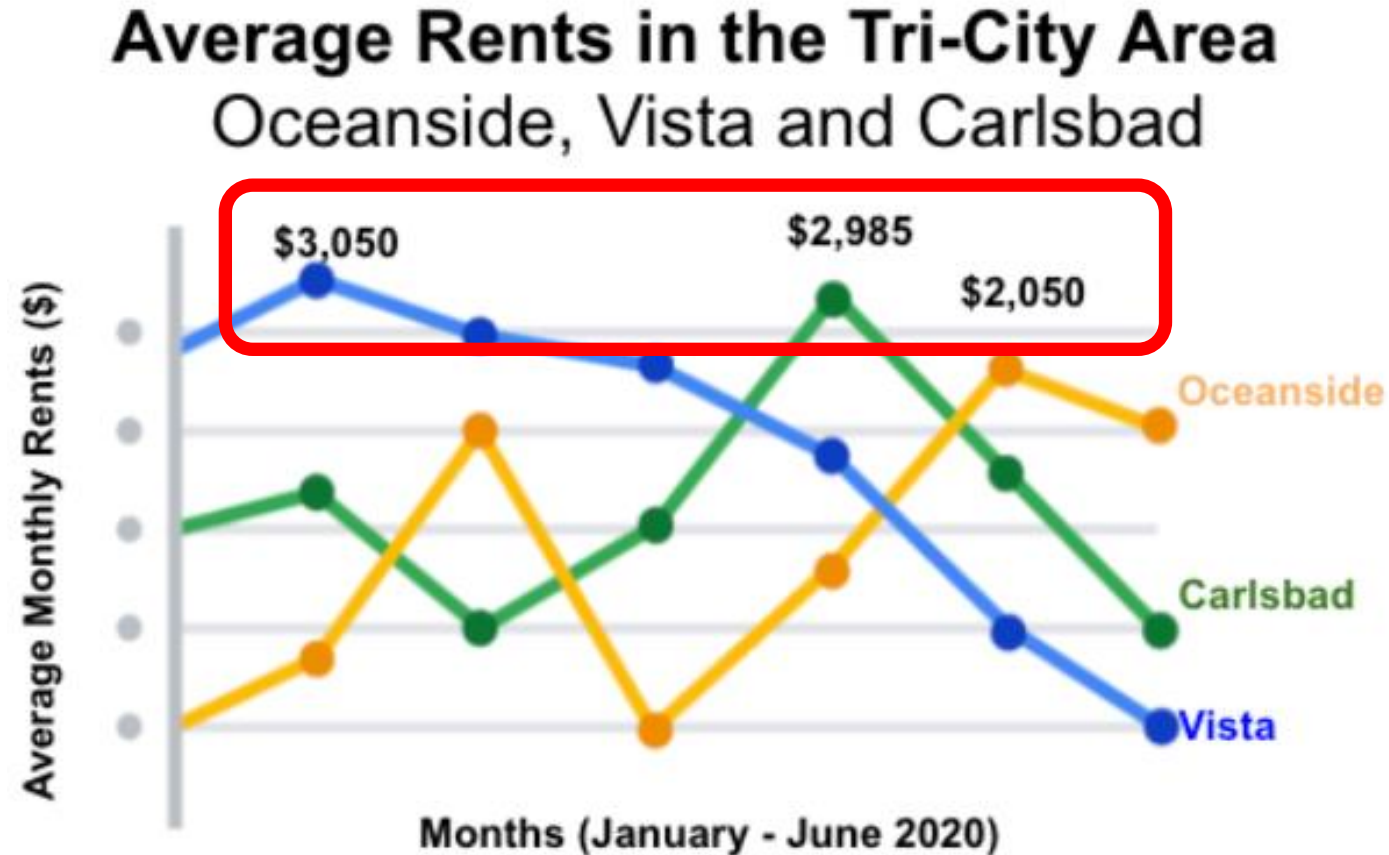
Effective Visual Cues

- Labels that identify



Effective Visual Cues

- Annotations that focus



5. Libraries

Python offers a variety of libraries with diverse functionalities for illustrating data. Each of these libraries has unique features and supports a range of graph types.

- Matplotlib

- Seaborn

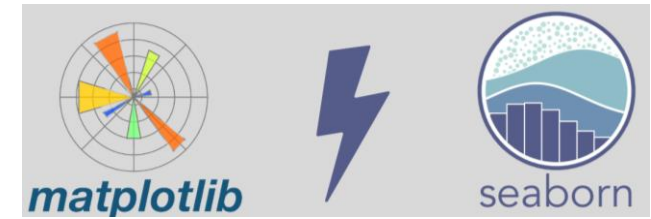
- Bokeh

- Plotly

Note: You may follow their full tutorial if you want to understand more details about these libraries.

PS: Provided that you are done with data cleaning and wrangling (i.e., you have a dataset you finally can call “clean”).

Matplotlib vs Seaborn

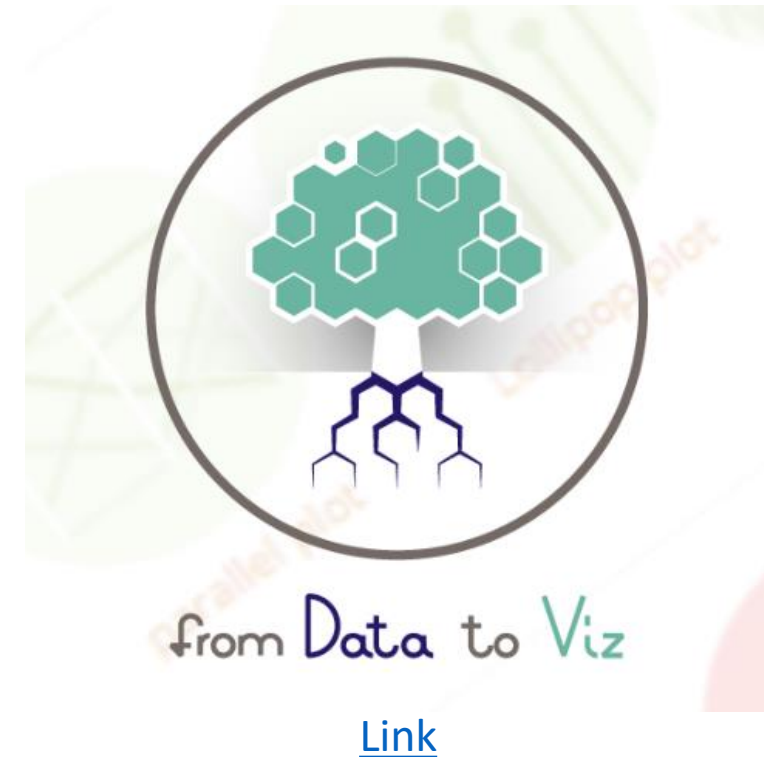


Characteristics	Matplotlib	Seaborn
Use Cases	Matplotlib plots various graphs using Pandas and Numpy	Seaborn is the extended version of Matplotlib which uses Matplotlib along with Numpy and Pandas for plotting graphs
Complexity of Syntax	It uses comparatively complex and lengthy syntax.	It uses comparatively simple syntax which is easier to learn and understand.
Multiple figures	Matplotlib has multiple figures can be opened	Seaborn automates the creation of multiple figures which sometimes leads to out of memory issues
Flexibility	Matplotlib is highly customizable and powerful.	Seaborn avoids a ton of boilerplate by providing default themes which are commonly used.

Matplotlib vs Seaborn vs Plotly

matplotlib	
PROS <ul style="list-style-type: none">• Versatile and accessible• Customizable• Good documentation• Universal data viz tool that plugs into many back ends	CONS <ul style="list-style-type: none">• Steep learning curve• Users need to know Python• Users need to understand the syntax of Matplotlib, which is based on the software, Matlab
seaborn	
PROS <ul style="list-style-type: none">• Quickly creates simple visualizations• Creates aesthetically pleasing visualizations	CONS <ul style="list-style-type: none">• Limitations on what you can customize• Visualizations are not as interactive• Users may need to simultaneously use Matplotlib
plotly	
PROS <ul style="list-style-type: none">• Accessible• Creates aesthetically pleasing visualizations• Easy to create interactive features within visualization	CONS <ul style="list-style-type: none">• Steep learning curve• Users need to know Python• Uses its own syntax

COURSE REPORT + **LIGHTHOUSE LABS**



Source: [Intro to 3 Data Viz Tools: Matplotlib, Seaborn, and Plotly](#)

6. Examples

What can be done...

- Please open the file “L7_work”.

7. Visualization Practicum

What you will be doing...

- It's group work. You will be assigned to groups.
- Prepare visualizations for a given dataset in file “gapminder_full.csv”, and finish by 2:40 pm.
- Present your visualizations / story briefly (3 min. per group).

You can get some insight from browsing the web, for example:

- <https://www.kaggle.com/code/tklimonova/gapminder-graph-using-python/notebook>;
- <https://www.data-to-viz.com/>

Example Dataset

Gapminder









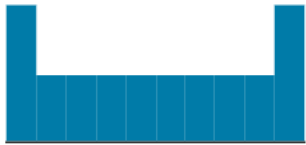

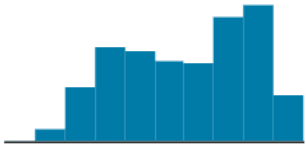

- This dataset contains data from the Gapminder.org website.

gapminder_full.csv (83.79 kB)

Detail Compact Column

About this file

This file 'gapminder_full.csv' contains the data from Gapminder website for 1952 till 2007.

 country 	# year 	# population 	 continent 	# life_exp 	# gdp_cap 
country of the world	year	amount of people that live in the country	continent	average time person is expected to live	gross domestic product divided by its total population
142 unique values	 19522007	 60.0k1.32b	Africa 37% Asia 23% Other (684) 40%	 23.682.6	 241114k

Yet, REMEMBER!

Reality will differ...



11. In-class assignment