Data visualization practice

Ewha GSIS Computational Social Science Workshop

Iegor Vyshnevskyi

Woosong University

October 14th, 2023

Agenda

1. Introduction to Data Visualization
2. Importance of Data Visualization
3. Chart Types
4. Visual Cues
5. Benefits of using R for Visualization
6. Intro to ggplot2 package in R

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

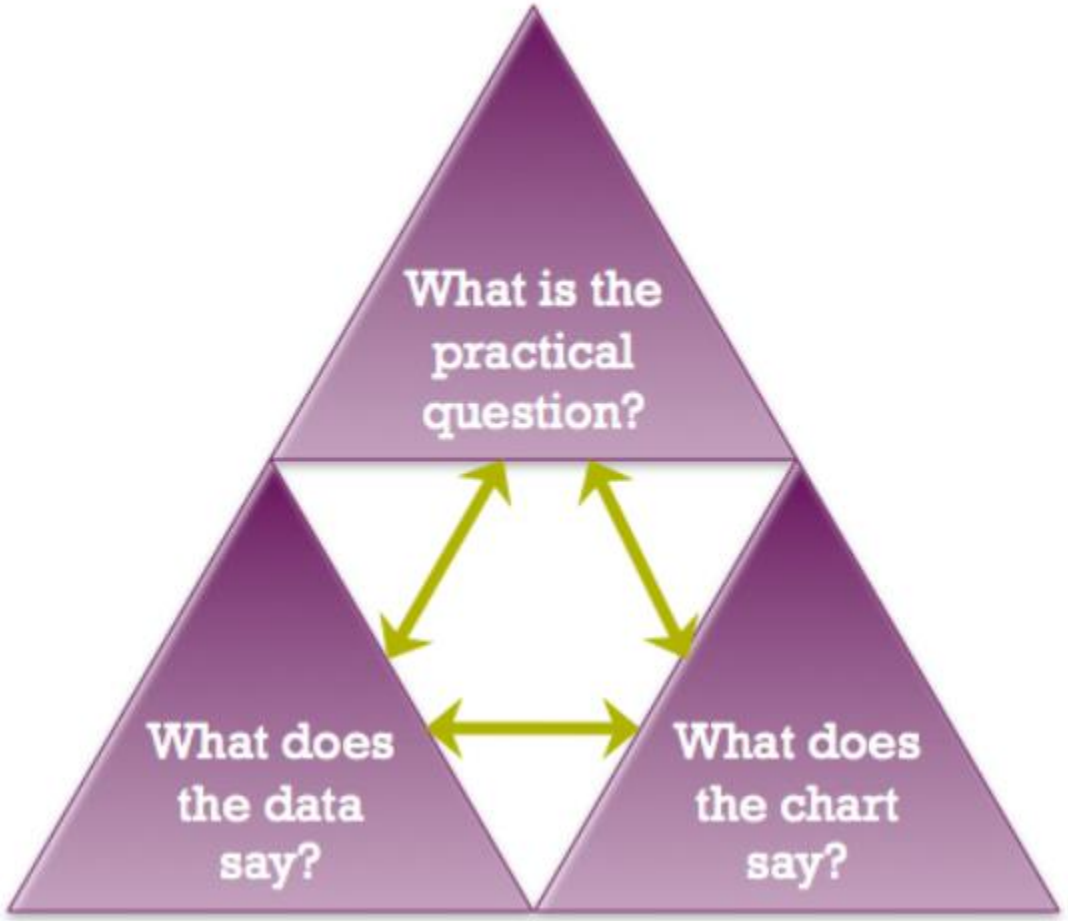
under the David McCandless method

Source: www.informationisbeautiful.net

- **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



Source: https://junkcharts.typepad.com/junk_charts/junk-charts-trifecta-checkup-the-definitive-guide.html

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.

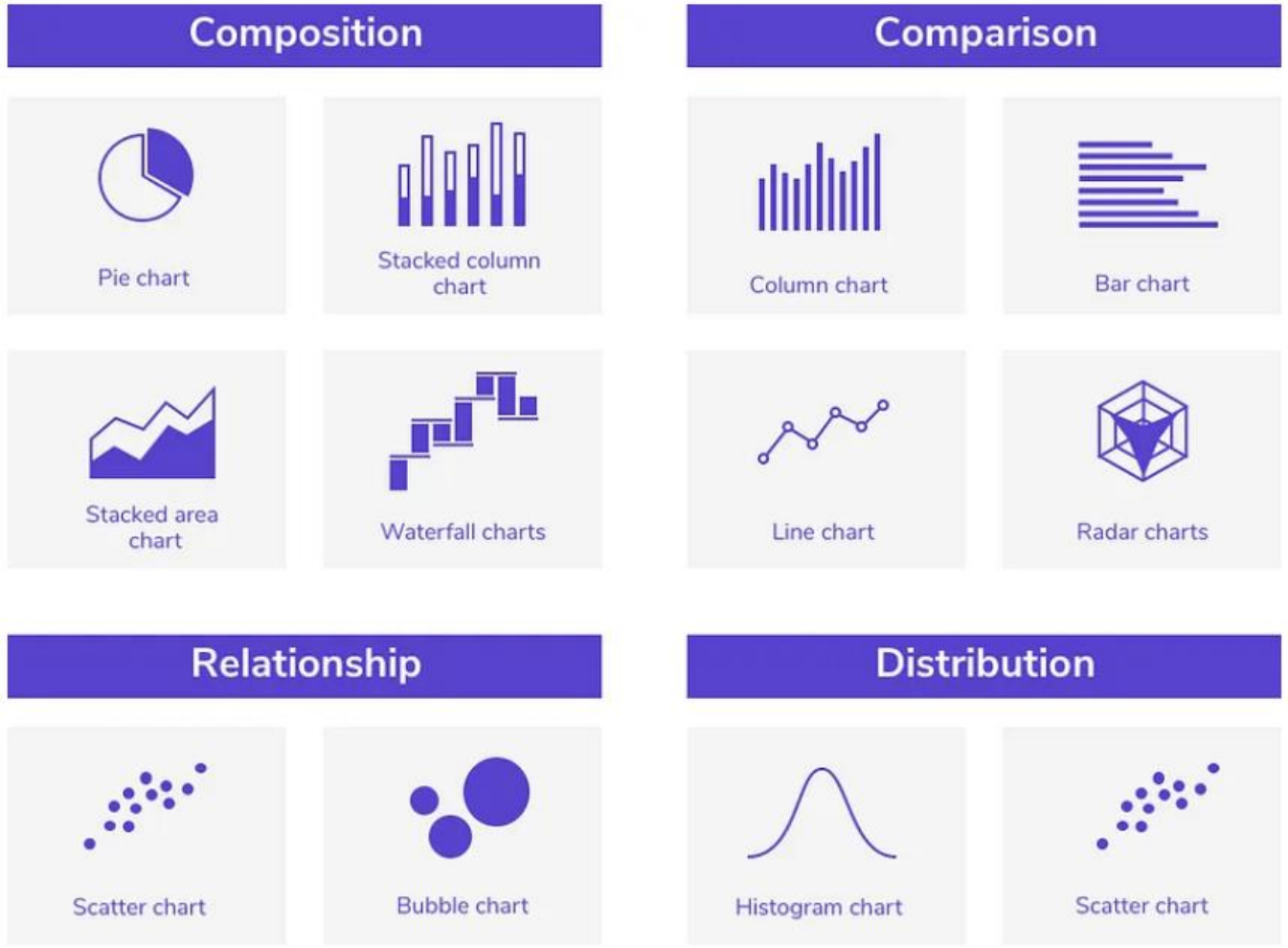
Source: Coursera course "Share Data Through Art of Visualization"

3. Chart Types

- Different types of visualization 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/*](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

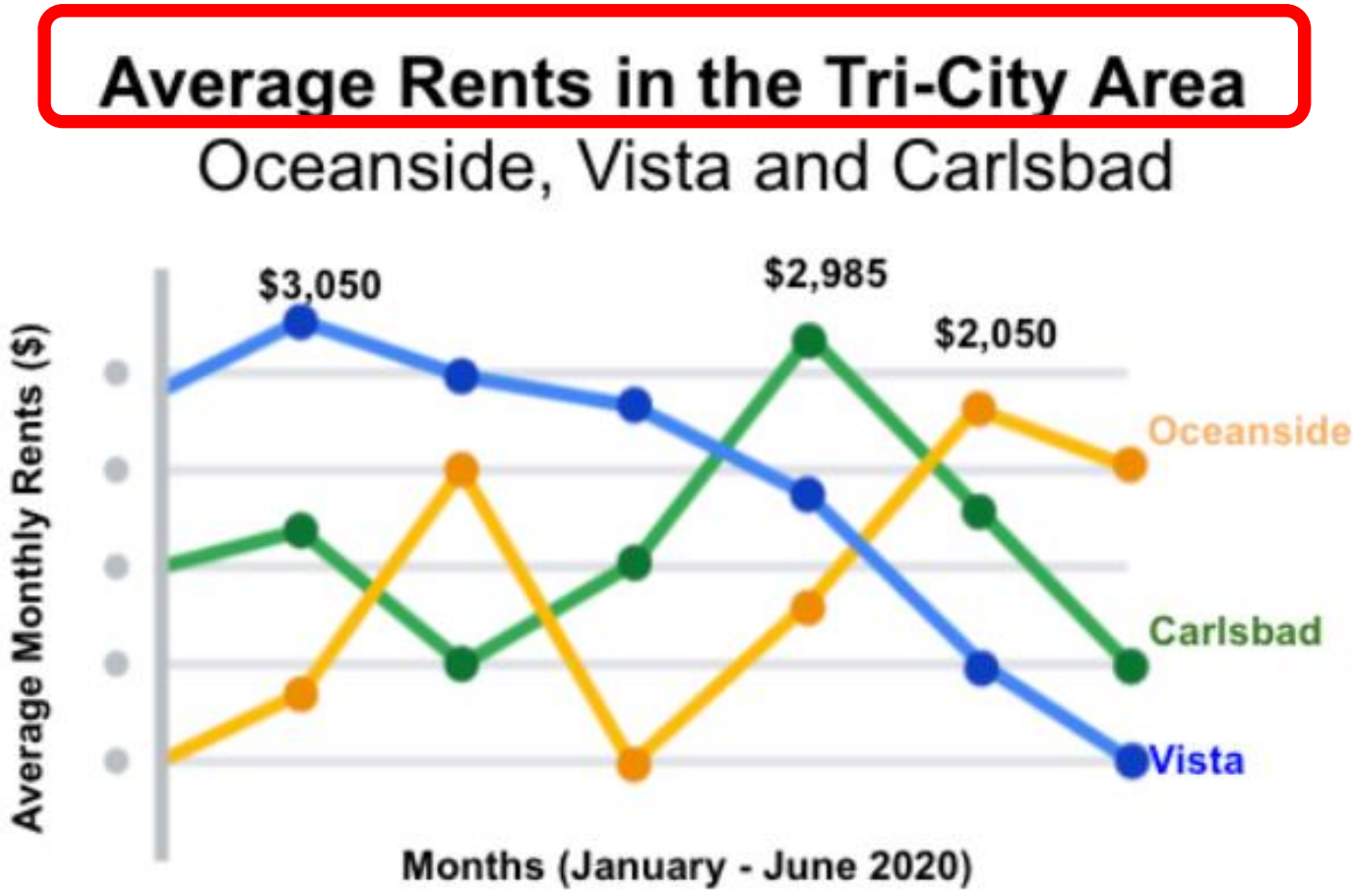
If you want to invite your audience into your presentation and keep them engaged, you have only ***5 seconds*** to catch their interest.

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

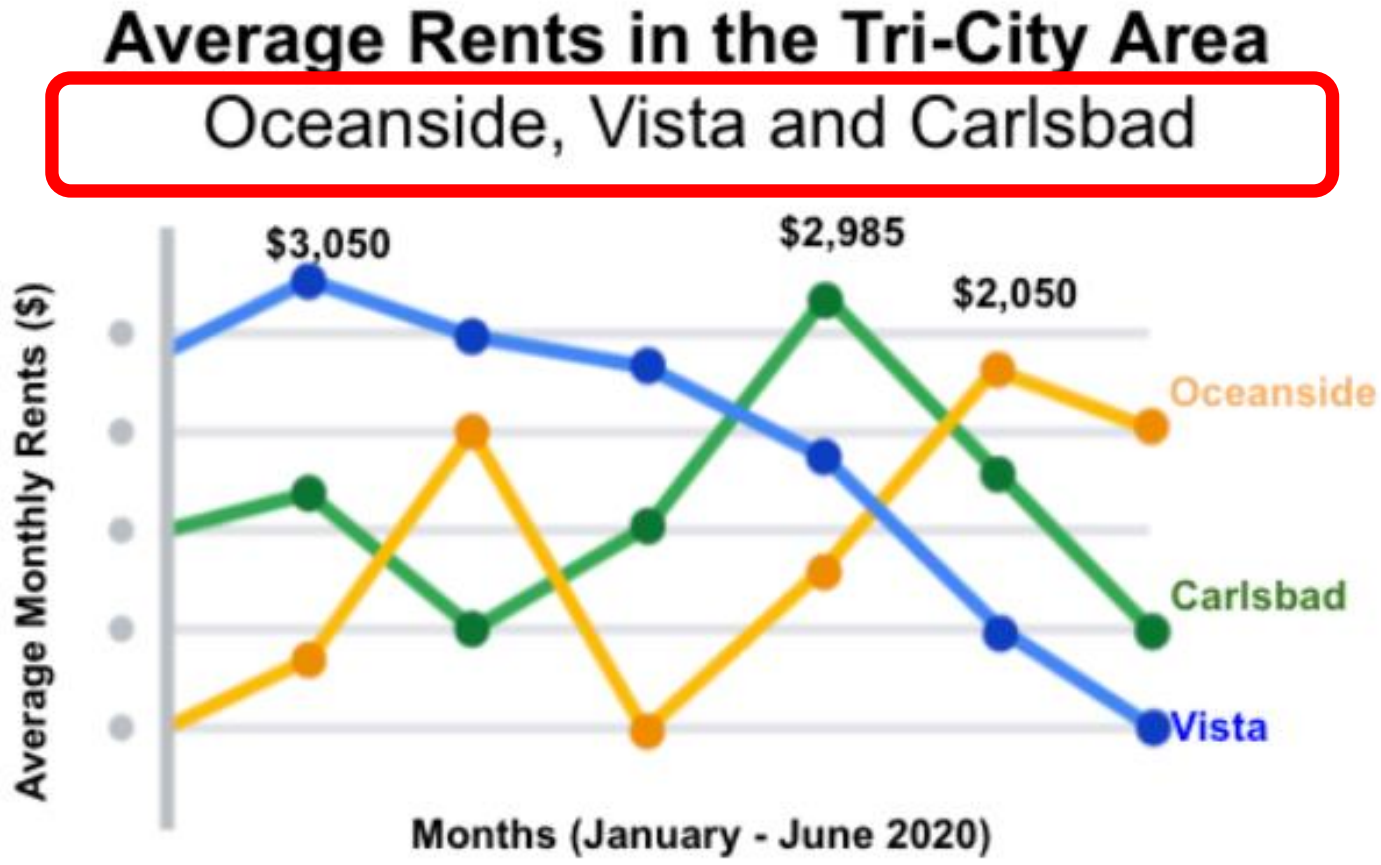
- **Headlines that pop**



Source: Coursera course "Share Data Through Art of Visualization"

Effective Visual Cues

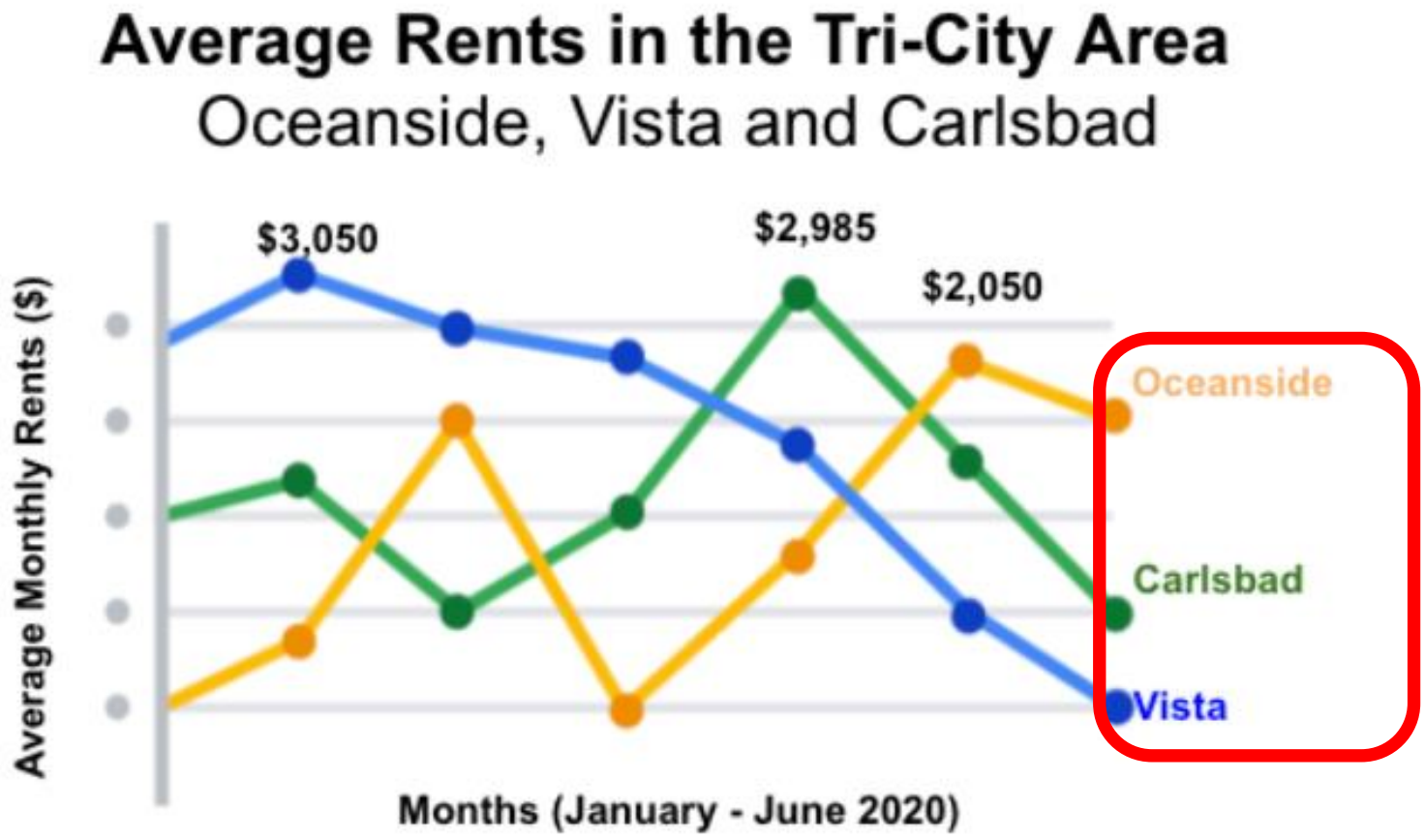
- **Subtitles that clarify**



Source: Coursera course "Share Data Through Art of Visualization"

Effective Visual Cues

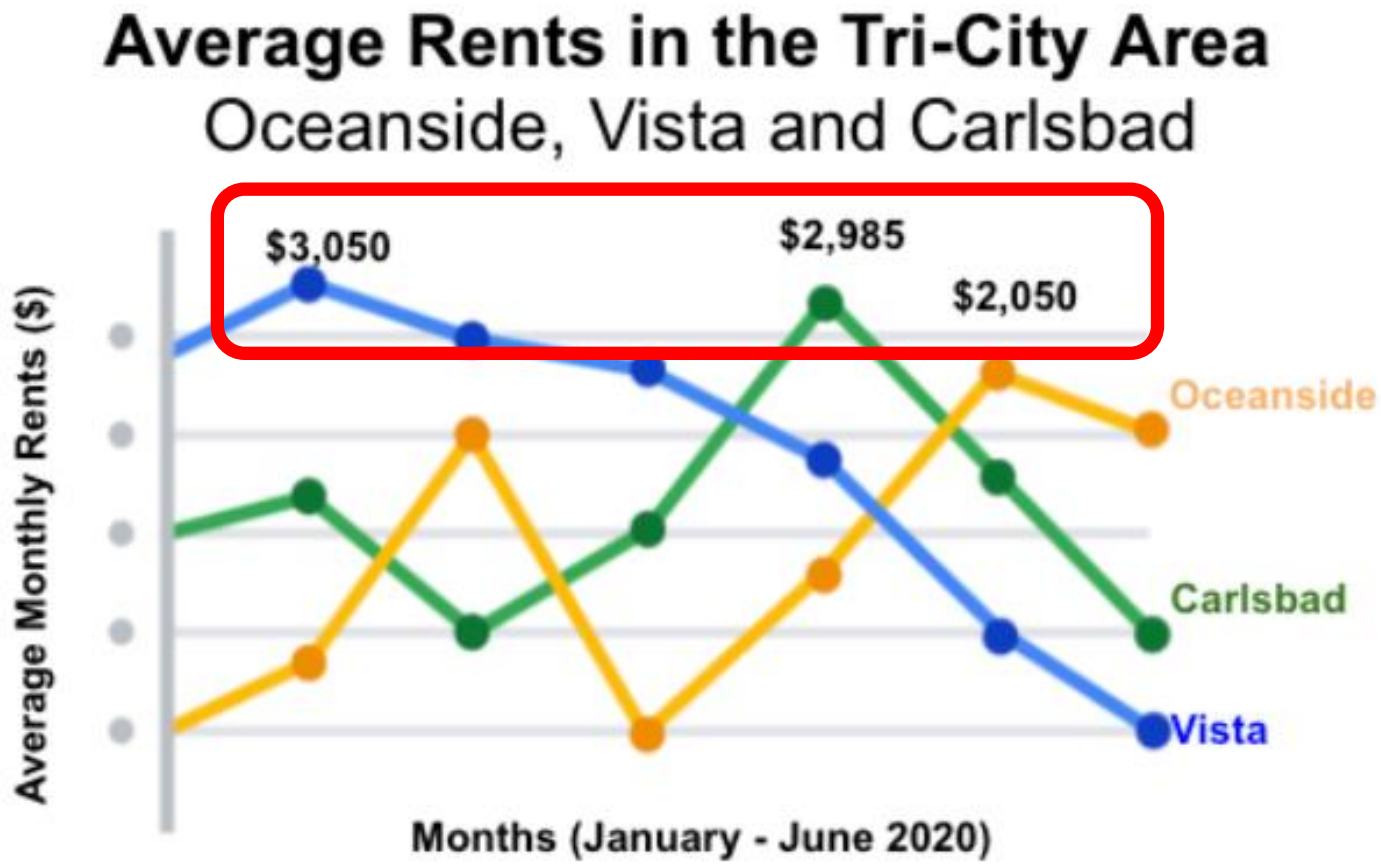
- **Labels that identify**



Source: Coursera course "Share Data Through Art of Visualization"

Effective Visual Cues

- **Annotations that focus**



Source: Coursera course "Share Data Through Art of Visualization"

5. Benefits of using R for Visualization

Key Benefits of using R for Visualization:

- ***Integration with Data Analysis:*** R has native support for working with data frames and matrices, allowing for seamless integration between analysis and visualization.
- ***Rich and Extensive Visualization Capabilities:*** R has a vast library of visualization packages, providing a wide range of chart types and customization options for static and interactive visualizations.
- ***Open-Source and Free:*** R is an open-source language that is free to download and use, making it a cost-effective solution for data visualization and analysis.

Key Benefits of using R for Visualization:

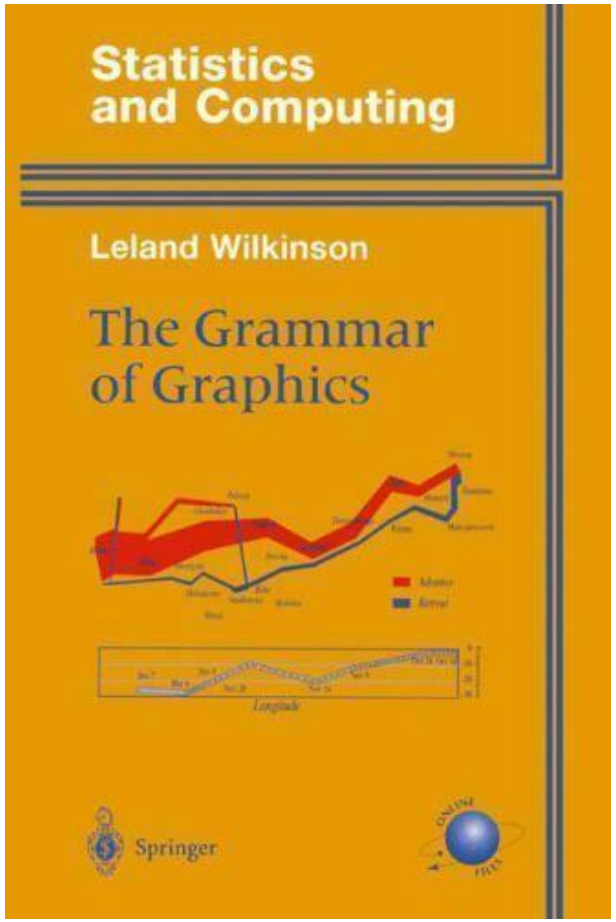
- ***Reproducibility and Reusability:*** R scripts can be used to create and save visualizations, allowing for easy reproduction and sharing of results. R code and packages are also widely available online, allowing for easy reuse and customization of existing visualization templates.
- ***Integration with Other Tools:*** R can be integrated with other tools and languages commonly used in data analysis, such as Python and SQL, allowing for a seamless workflow across different stages of data analysis and visualization.

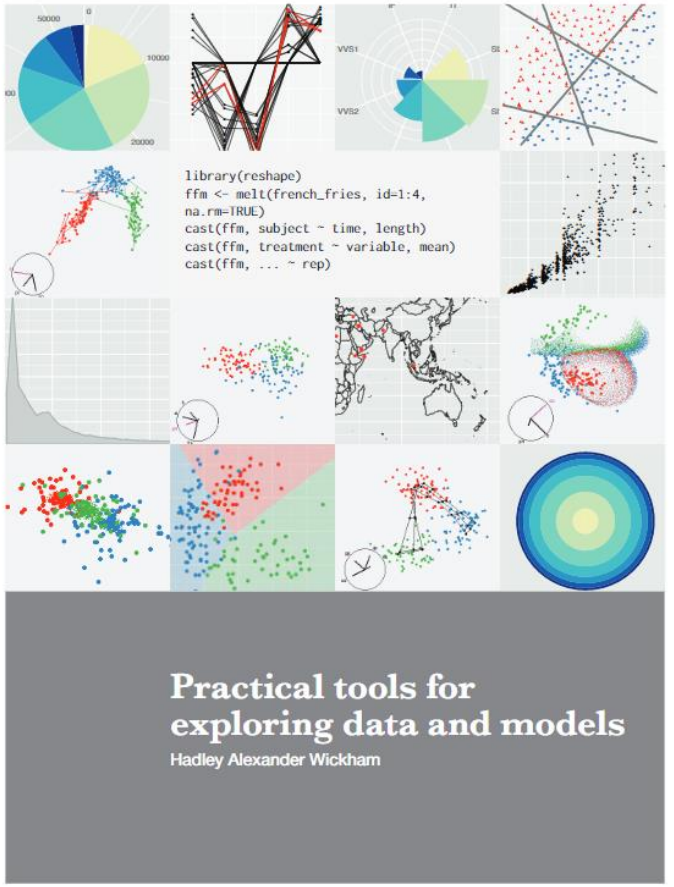
When R is more reasonable for data visualization

- ***Complex Data Manipulation:*** R provides greater control over data cleaning and transformation, making it better suited for large or messy datasets.
- ***Customization and Control:*** R provides more customization and control over visualizations, allowing for highly customized or specialized visualizations.
- ***Statistical Analysis:*** R is better suited for advanced statistical analysis and modeling, making it useful for visualizing and communicating results to stakeholders.
- ***Programming Flexibility:*** R provides more flexibility and customization options than Tableau's point-and-click interface, making it easier to create complex or customized visualizations.
- ***Cost and Licensing:*** R is an open-source language that is free to use, making it a more cost-effective solution for data visualization and analysis.

6. Intro to ggplot2 package in R

The Beginnings of ggplot2

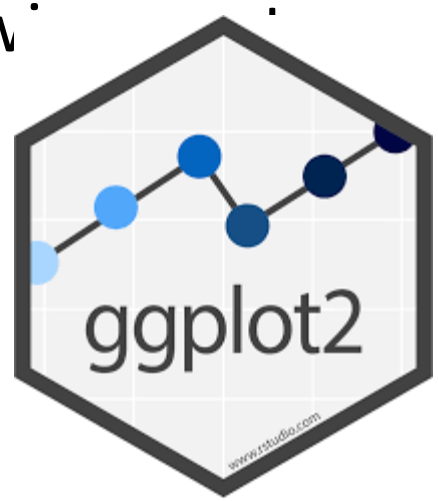




ggplot2 is an R package, a part of *tidyverse*, which provides a large variety of plotting functionality to enable better and highly customizable graphs.

The basic structure of a ***ggplot2*** code involves creating a plot object and adding layers to it, such as data points, lines, labels, and axes.

ggplot2 allows for a high degree of customization, allow
control almost every aspect of the plot.



When it is better to use basic plot functions

- For quick and simple visualizations that don't require a lot of customization.
- When working with small datasets that don't require advanced customization or layering.
- If you're already familiar with basic plot functions and want to quickly create a visualization without learning a new syntax or package.
- For creating simple charts such as histograms or bar charts.
- For scatter plots with large datasets as ggplot may slow down.

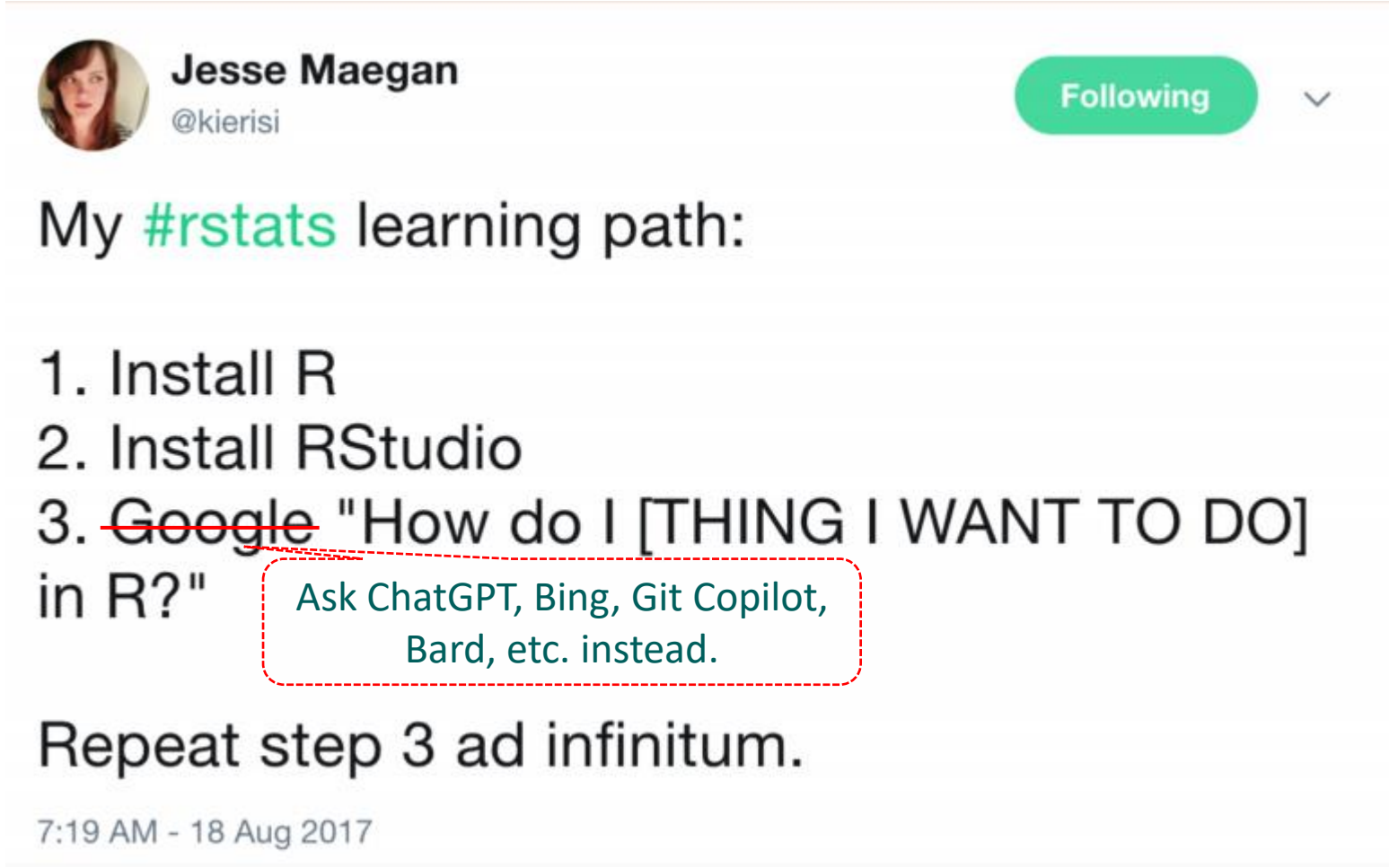
When it is better to use ggplot2

- For creating complex visualizations with multiple layers and aesthetics.
- When you want to customize the plot in detail.
- For creating more specialized plot types.
- When working with large datasets and need to use facets to split the plot.
- For adding advanced statistical methods, such as smoothing lines or correlation coefficients.
- For creating high-quality graphics suitable for publication or presentation purposes.

Some readings

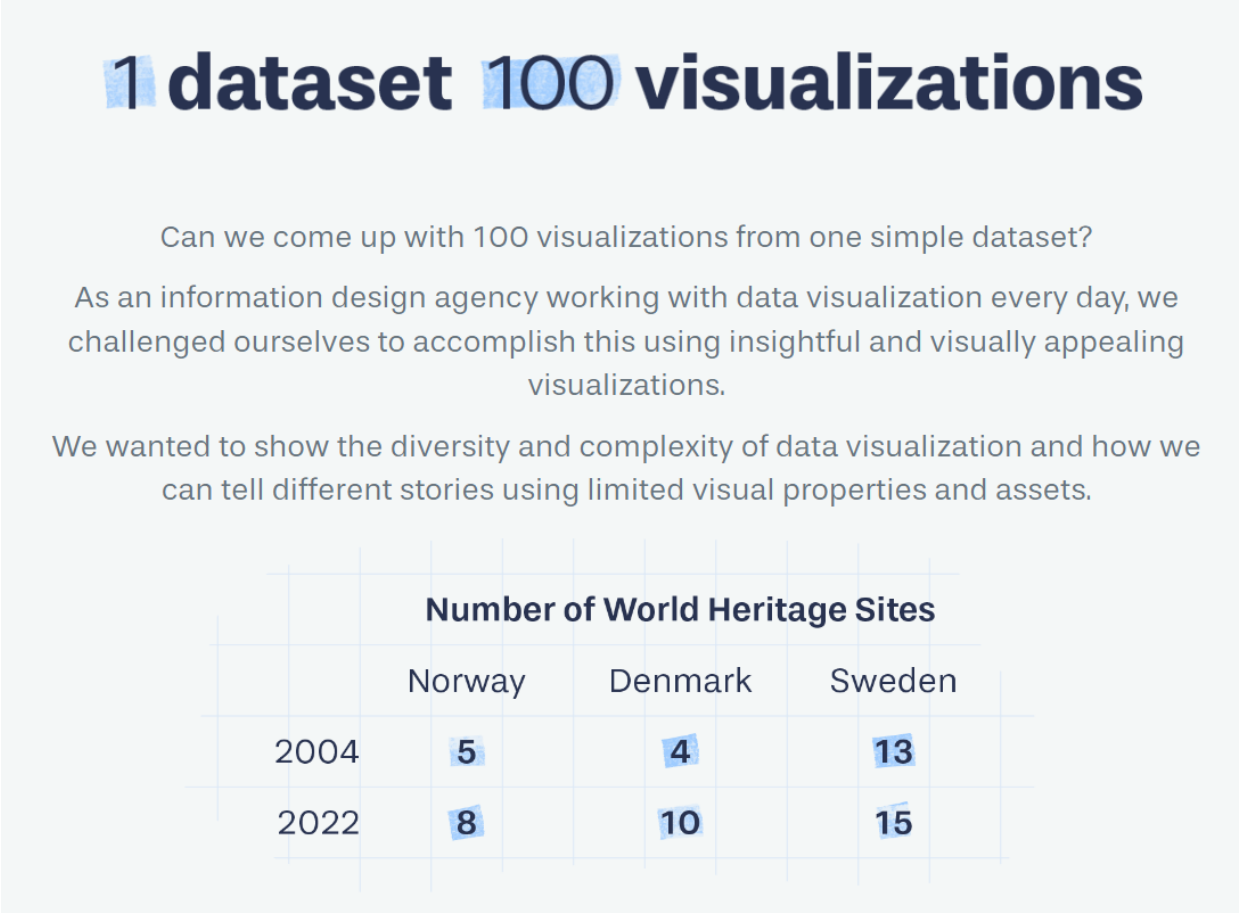
- Franconeri et al. (2021). The Science of Visual Data Communication: What Works.
- Kieran Healy (2019). Data Visualization.
- Claus Wilke (2019). Fundamentals of Data Visualization.
- Hadley Wickham et al. ggplot2: Elegant Graphics for Data Analysis (3ed.).
- Garrick Aden-Buie. A Gentle Guide to the Grammar of Graphics with ggplot2.

How to work with R/ggplot2



If you think that there is nothing you can show with this data.

Recall this



Source: https://100.datavizproject.com/

It takes time…

So be patient ☺

ggplot2 practical use

Let's have some fun!