

Data Visualization: Introduction and Overview

```
$ echo "Data Science Institute"
```

Acknowledgement

We wish to acknowledge this land on which the University of Toronto operates. For thousands of years it has been the traditional land of the Huron-Wendat, the Seneca, and most recently, the Mississaugas of the Credit River. Today, this meeting place is still the home to many Indigenous people from across Turtle Island and we are grateful to have the opportunity to work on this land.

Welcome!

Introductions:

- TF: Eshta Bhardwaj (she/her) eshta.bhardwaj@mail.utoronto.ca
- LS: Fan fw2400@cumc.columbia.edu
- LS: Kasra k.vakiloroayaei@utoronto.ca

Your introductions?

Today, we will...

- Go through slide deck #1: intro and overview
- Admin / housekeeping
- Discuss assignment 1
- Go through slide deck #2: getting started with matplotlib

Overview of this slide deck

- Explore why we should care about data visualization
- Question what makes 'good' data visualization
- Introduce a range of software and tools that are used for data visualization

Case Study: Why should we care about data visualization?

- Vision is a powerful communication channel
- Vision can easily pick up interesting patterns
 - Pre-attentive processing is faster than eye movement
- Just the data is not always enough
 - And summaries lose information

Anscombe's Quartet

| | x1 | x2 | x3 | x4 | y1 | y2 | y3 | y4 |
|---|----|----|----|----|------|------|-------|------|
| 1 | 10 | 10 | 10 | 8 | 8.04 | 9.14 | 7.46 | 6.58 |
| 2 | 8 | 8 | 8 | 8 | 6.95 | 8.14 | 6.77 | 5.76 |
| 3 | 13 | 13 | 13 | 8 | 7.58 | 8.74 | 12.74 | 7.71 |
| 4 | 9 | 9 | 9 | 8 | 8.81 | 8.77 | 7.11 | 8.84 |
| 5 | 11 | 11 | 11 | 8 | 8.33 | 9.26 | 7.81 | 8.47 |
| 6 | 14 | 14 | 14 | 8 | 9.96 | 8.10 | 8.84 | 7.04 |

Table 5.2: Mean and standard deviation for Anscombe's quartet

| Dataset | x mean | x sd | y mean | y sd |
|---------|--------|------|--------|------|
| 1 | 9 | 3.3 | 7.5 | 2 |
| 2 | 9 | 3.3 | 7.5 | 2 |
| 3 | 9 | 3.3 | 7.5 | 2 |
| 4 | 9 | 3.3 | 7.5 | 2 |

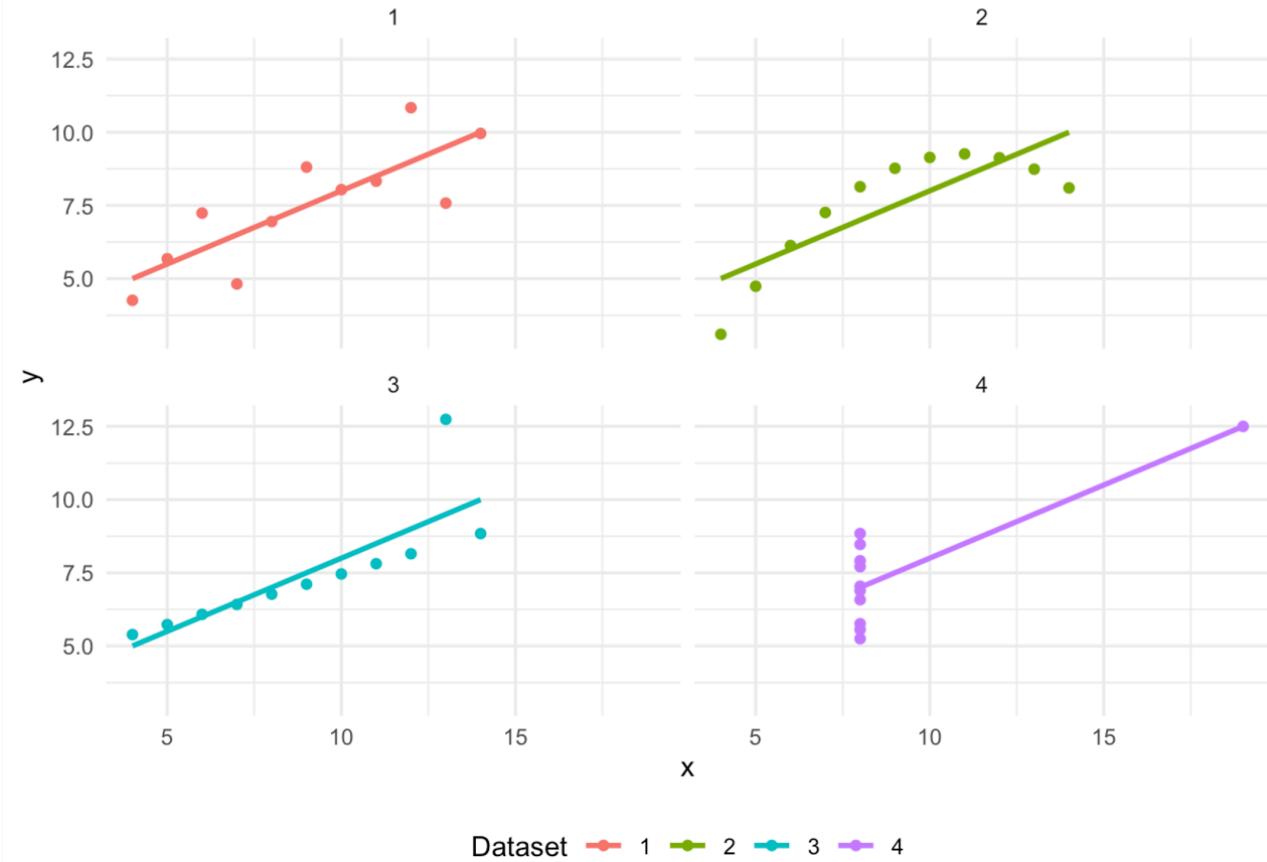
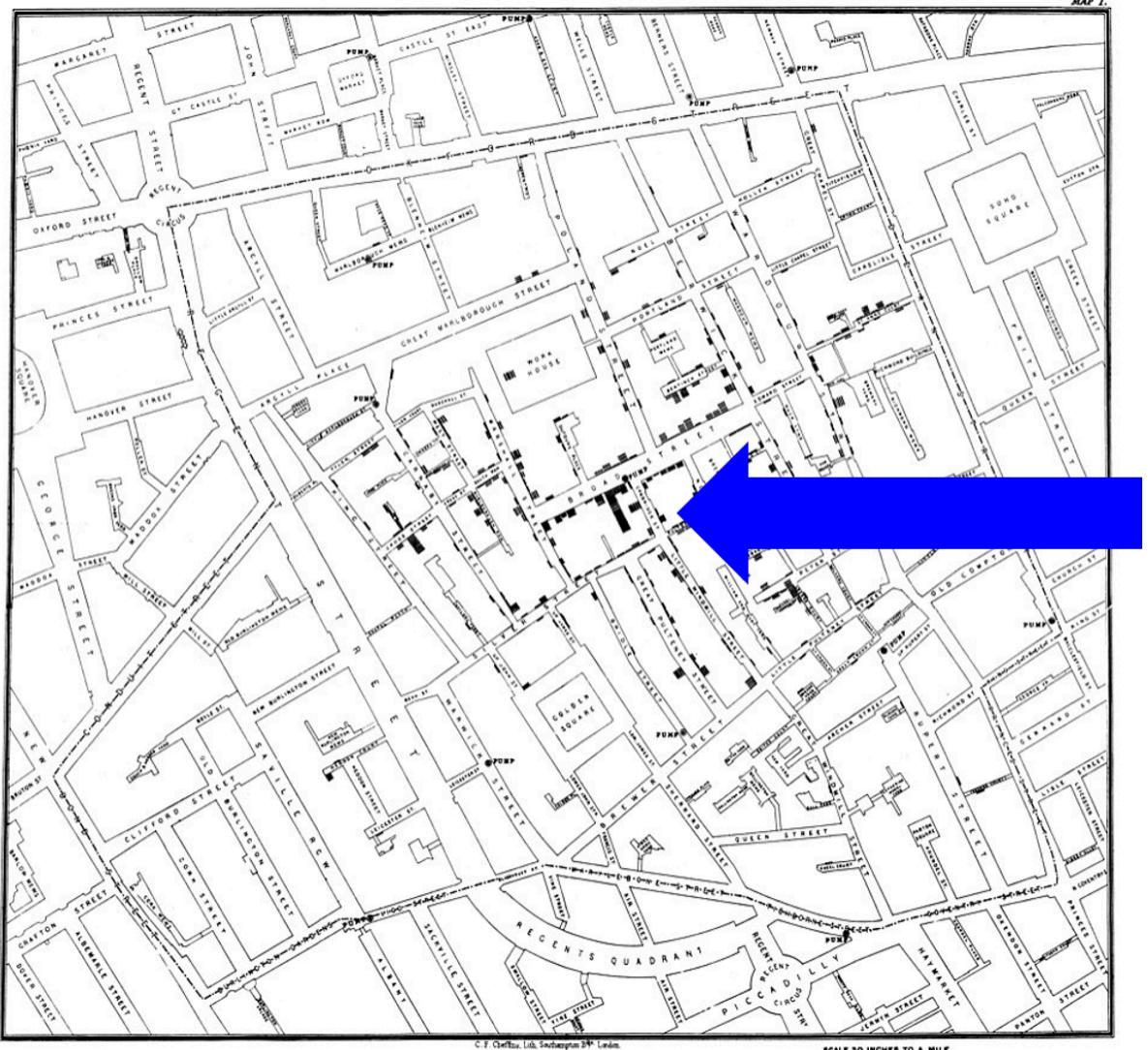
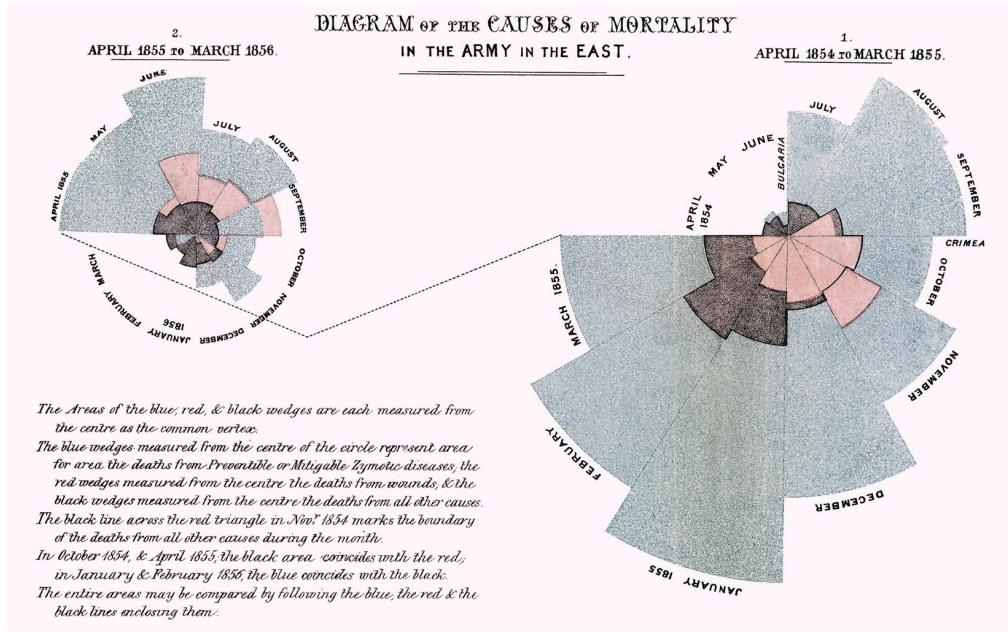


Figure 5.2: Recreation of Anscombe's Quartet

- No matter how good or groundbreaking our data science work is, if we can't communicate it, its impact will be severely limited.
- Often, the best way to communicate insights from data is in visual format.
- We can see examples of this idea throughout history.

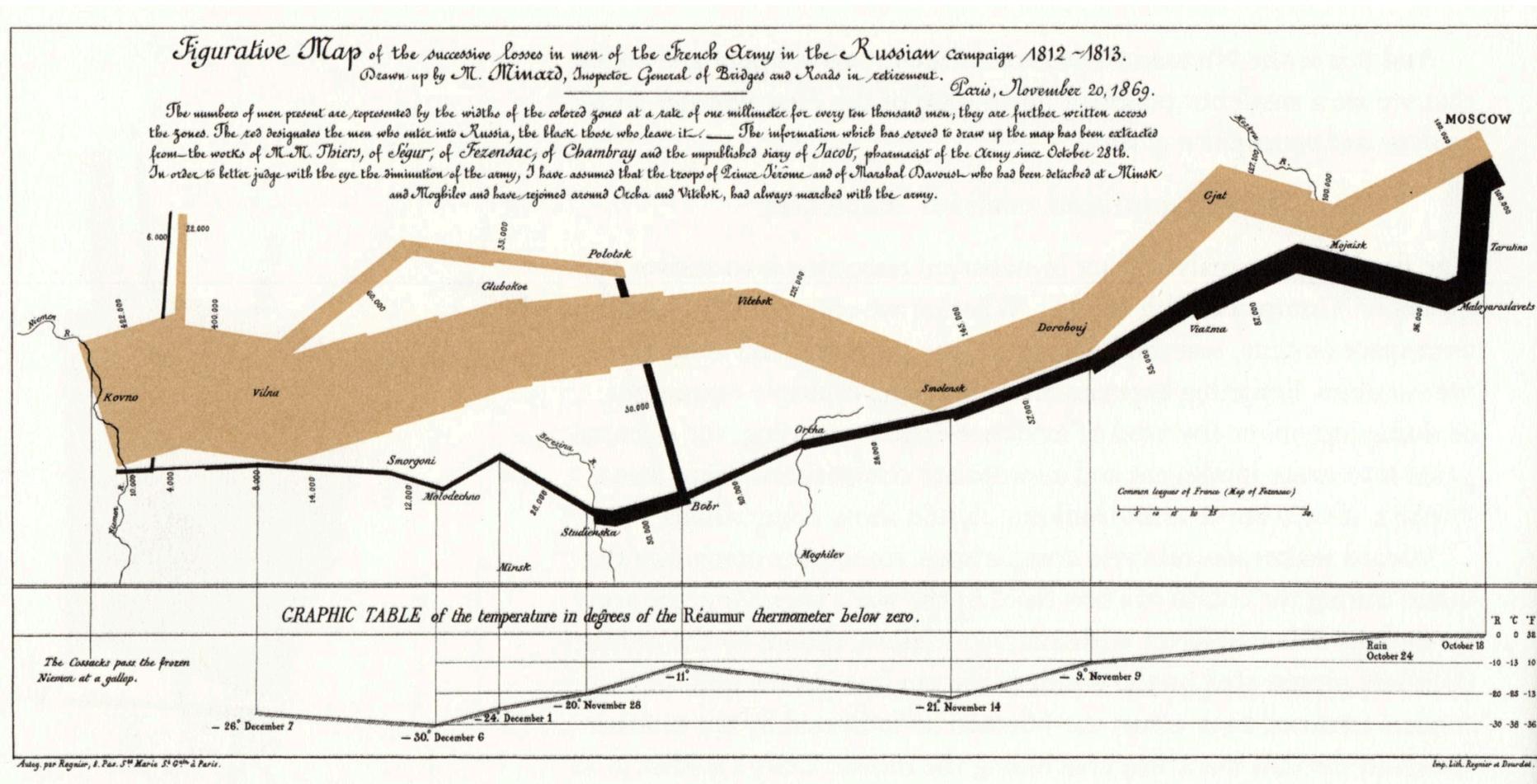


- By plotting black bars at the locations of each cholera death, Dr. John Snow was able to provide evidence [tracing an 1854 outbreak](#) to a specific water pump.
- This visualization is [often cited](#) as one of the early origins of the field of epidemiology.



- This 1858 visualization by Florence Nightingale shows soldiers' deaths due to wounds in battle (pink), other causes (black), and disease (blue) and was used to advocate for prioritizing nutrition, ventilation, and shelter, revolutionizing army medicine.

- Minard, in 1869, visualized: "Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812–1813."



So why visualize information?

- These historical examples gave us different reasons
 - Cholera outbreak → Allows us to see data in context
 - Nightingale's rose diagram → Presents an argument and advocates
 - Napoleon's Russian Campaign → An example of story telling

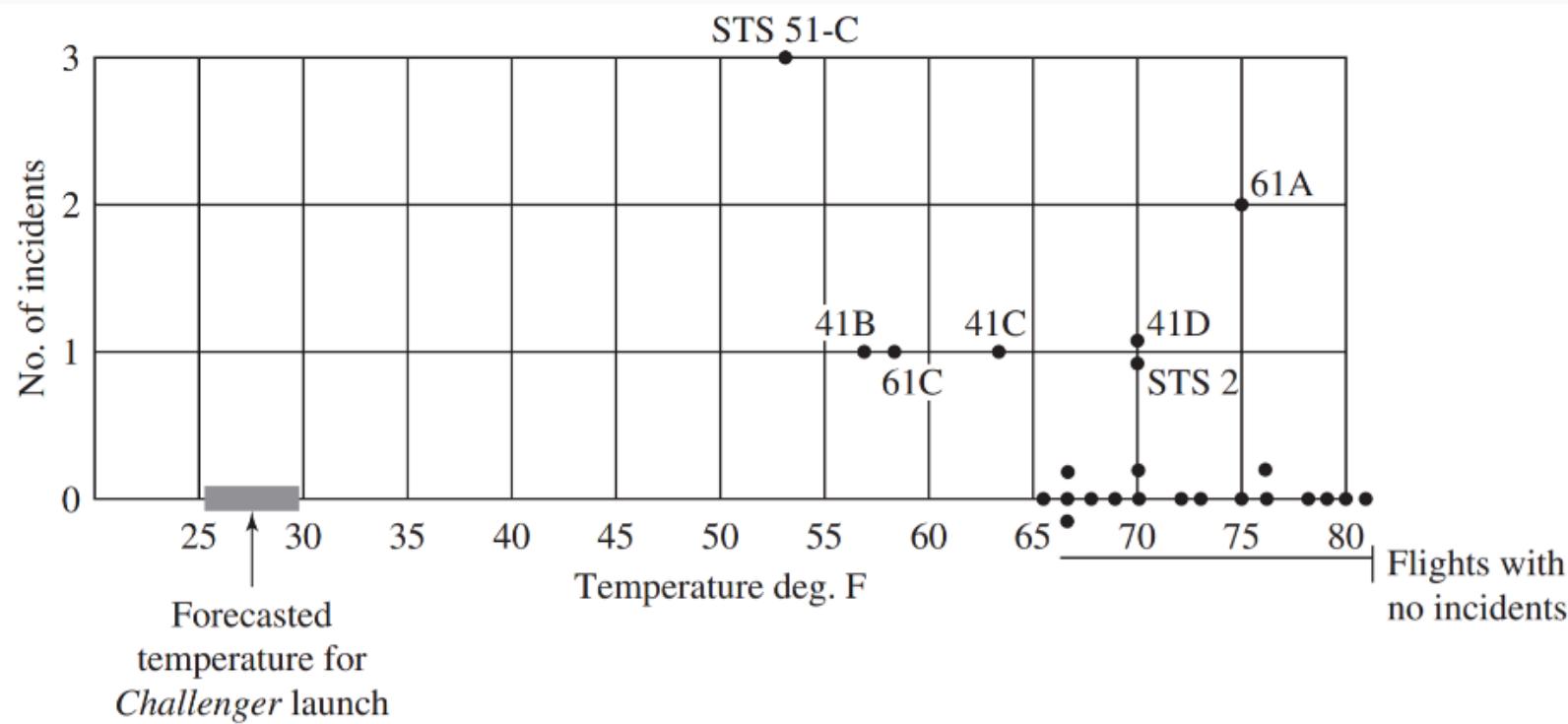
But data visualization is not always so straightforward...

The Challenger Disaster

- In 1986, the space shuttle *Challenger* exploded 73 seconds after launch, killing everyone aboard.
- The explosion occurred because hot propellant gases burned through rubber seals ("O-rings") on the shuttle's right solid rocket booster.
- For months, up until the night before the launch, concerns about the O-rings and the safety of the launch had been raised, but launch proceeded anyway

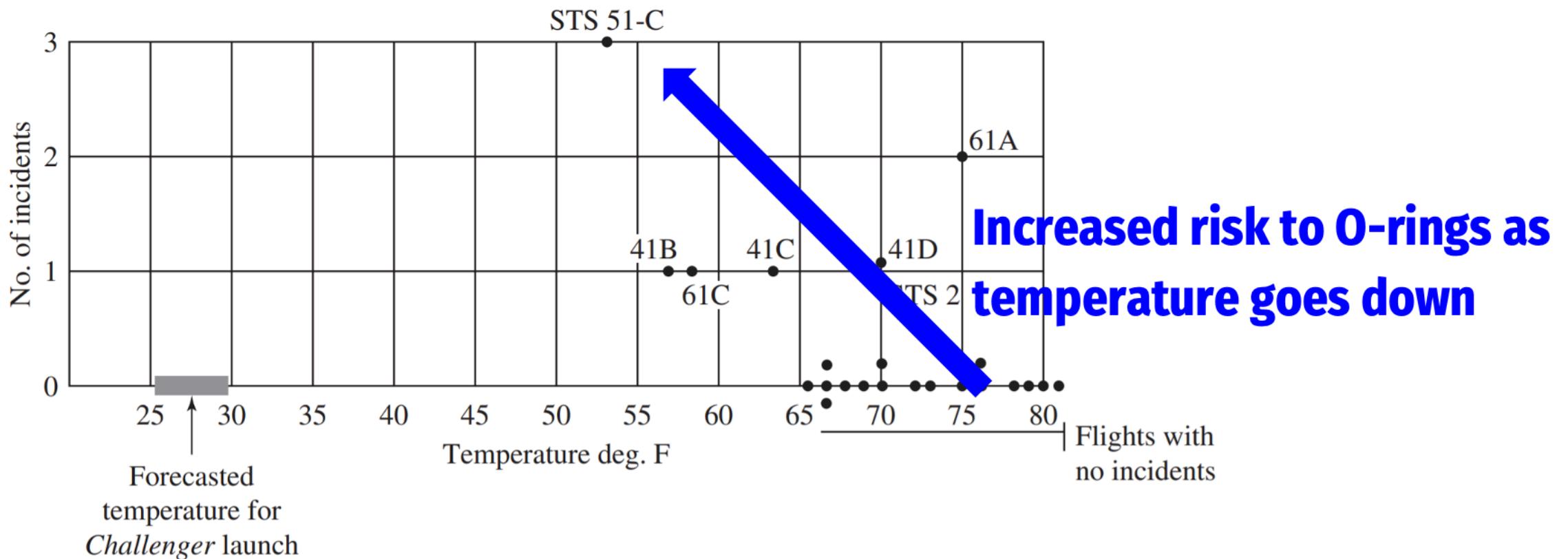
The Challenger Disaster - Tufte's Graph

- In 1997, Edward Tufte famously published a visualization showing the relationship between temperature during launches of test shuttle flights (pre-*Challenger*) and incidents of damage to O-rings (image: Tufte, 1997 



The Challenger Disaster - Tufte's Graph

- Tufte's graph seems to show that as launch temperatures decreased, O-ring incident rate increased; thus, by launching at temperatures lower than those tested, NASA unnecessarily endangered *Challenger* (image: Tufte, 1997 

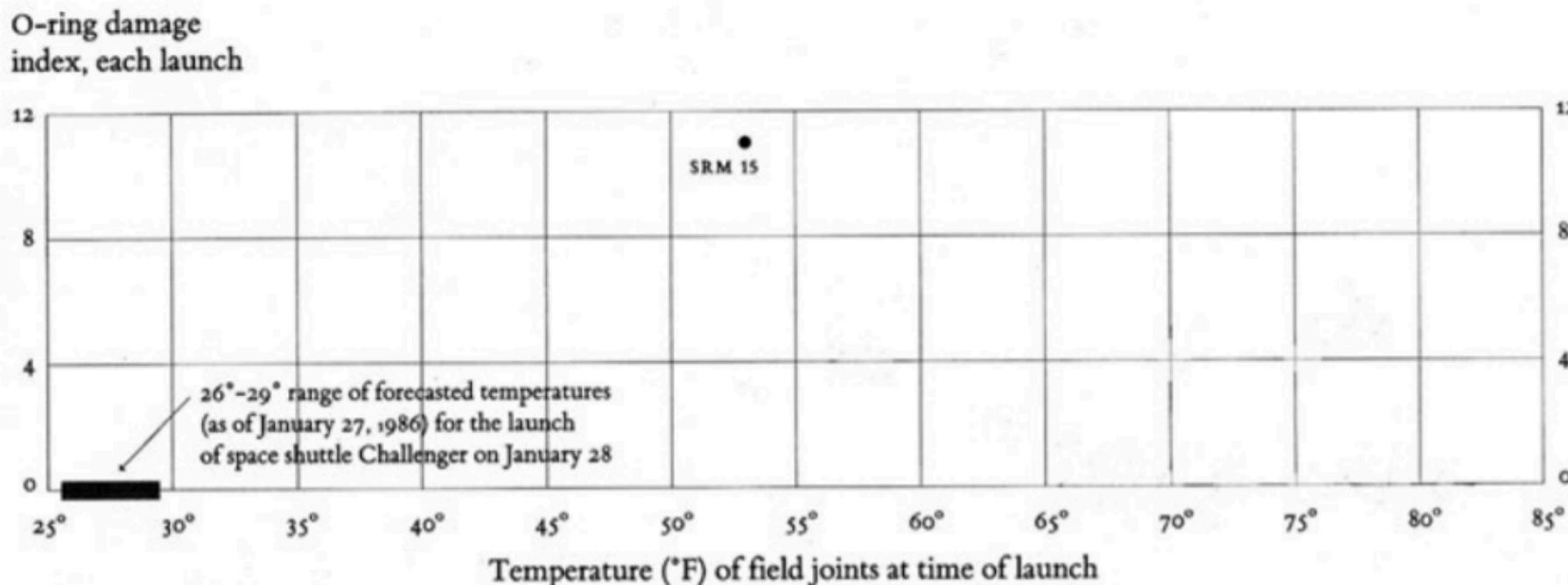


The Challenger Disaster - Tufte's Graph

- Tufte argued that the engineers responsible for communicating the results of the O-ring tests had failed to represent the data that might have saved the crew of *Challenger*.
- Tufte's graph is used as a classic case study demonstrating the importance of data visualization as a way to communicate complex information.
- Another case of data visualization saving lives?

The Challenger Disaster - But...

- The accuracy of Tufte's visualization has [been debated](#), with Robison identifying several errors in Tufte's use of data.
- Robison suggests that only one test launch had actually produced relevant O-ring temperature data, producing this visual instead: (image: Robison, 1997, 2002 



The Challenger Disaster - But...

- Per Tufte, the *Challenger* disaster is a case where good data visualization could have facilitated understanding of complex data and saved lives.
- Per Robison, Tufte's work is a case of bad data visualization unfairly placing blame and leading the audience to a faulty conclusion.

Other Challenges of Visualizing Information

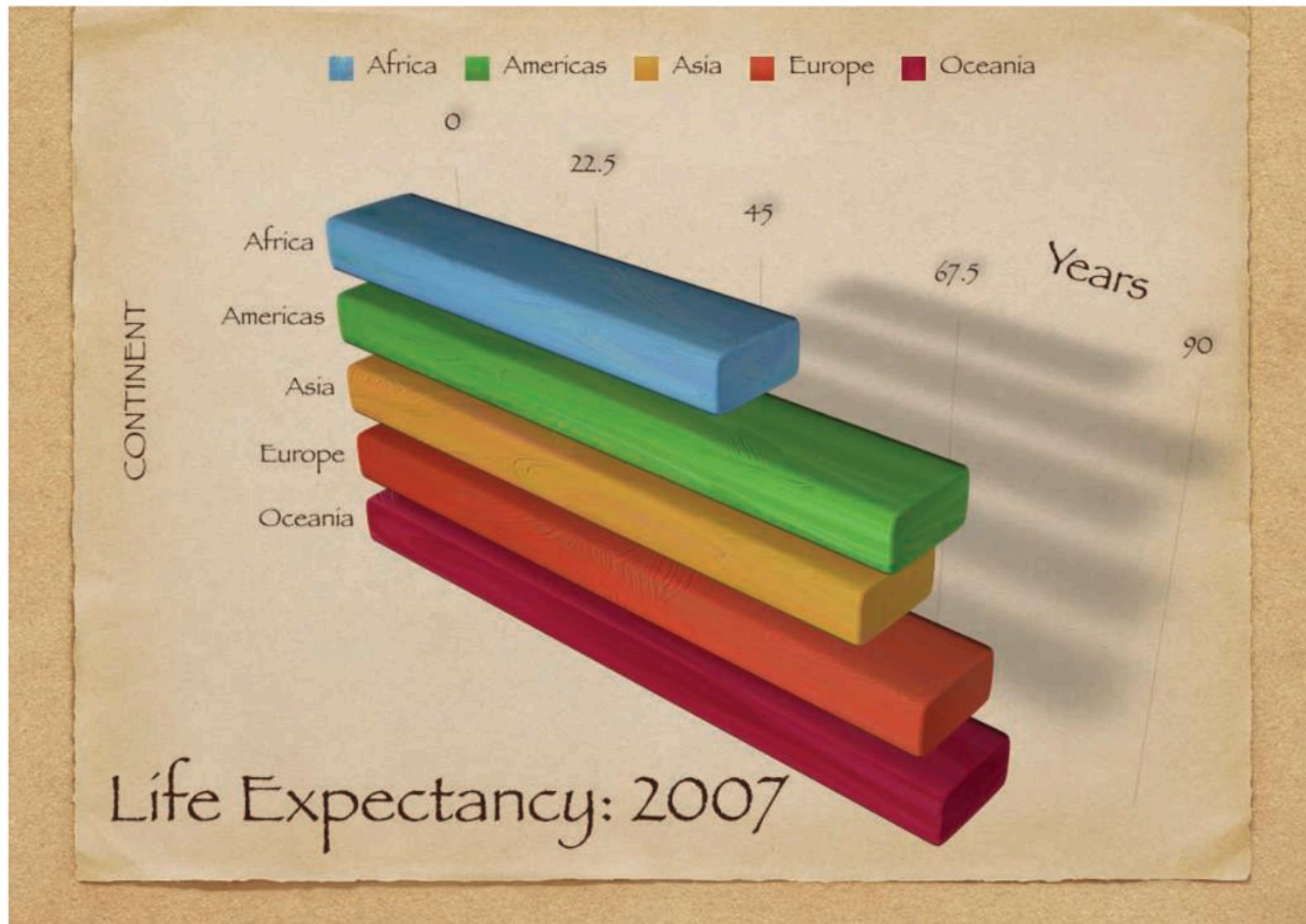
- Computational capacity
- Human perceptual and cognitive capacity
- Display capacity

The choices we make about visualizing our data have consequences - so how do we make better ones?

Activity: What is 'good' data visualization?

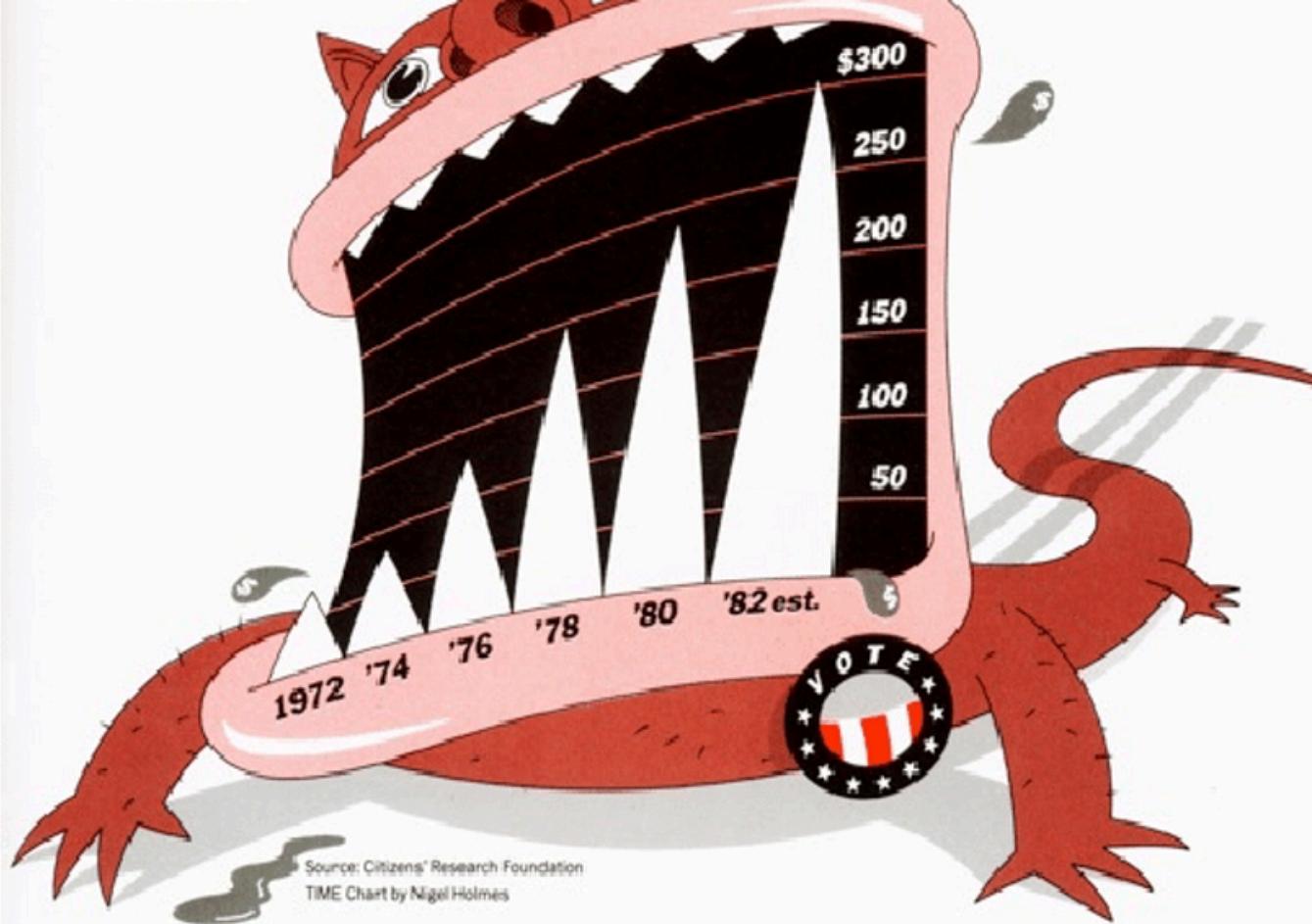
Activity

- Look at the following examples of data visualizations. For each, consider:
 - Is the visualization pleasing to look at?
 - Does the visualization accurately represent data?
 - Can we understand what message the maker of the visualization is attempting to convey?
 - Consider factors such as colour, size, use of images. Is this a 'good' data visualization? Why or why not?

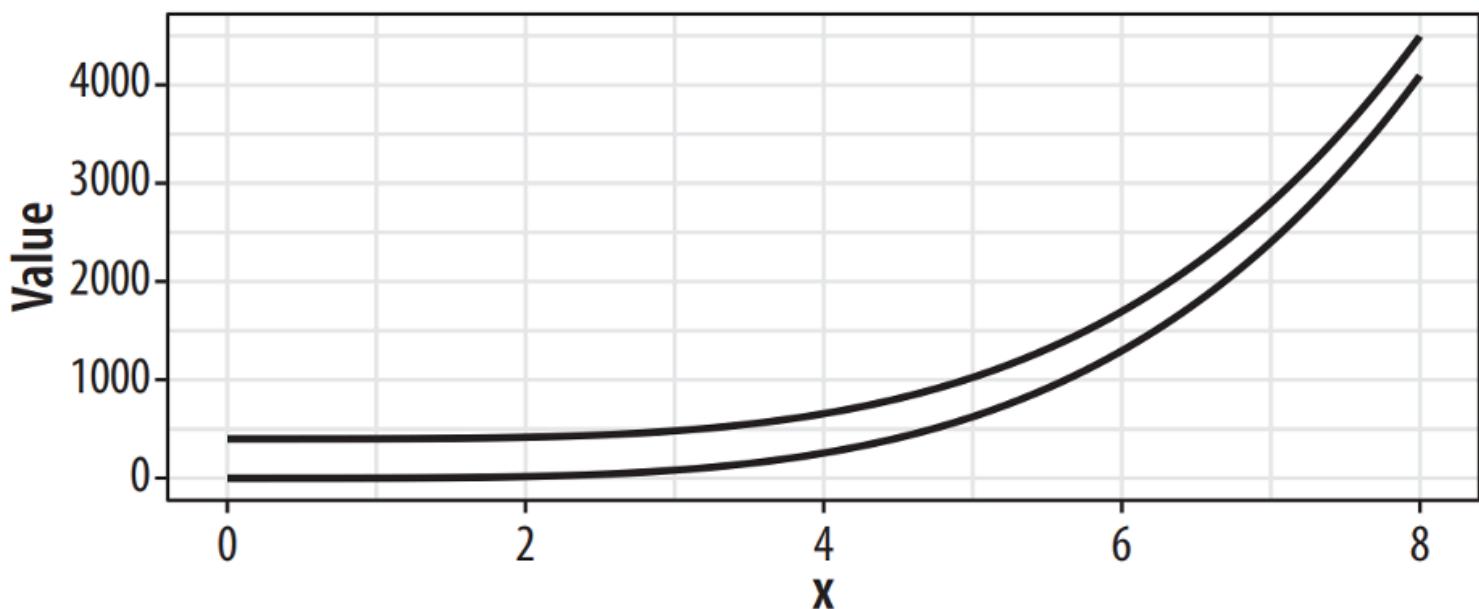
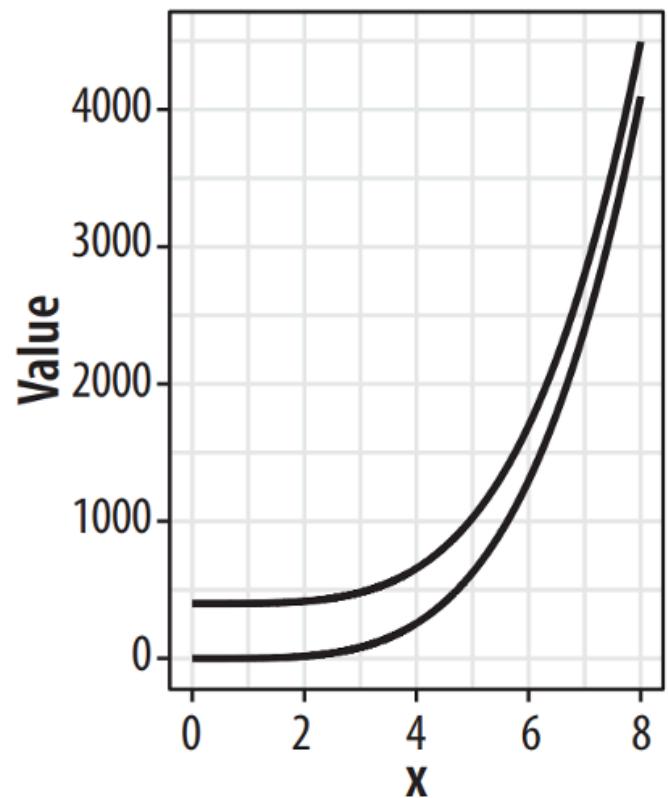


MONSTROUS COSTS

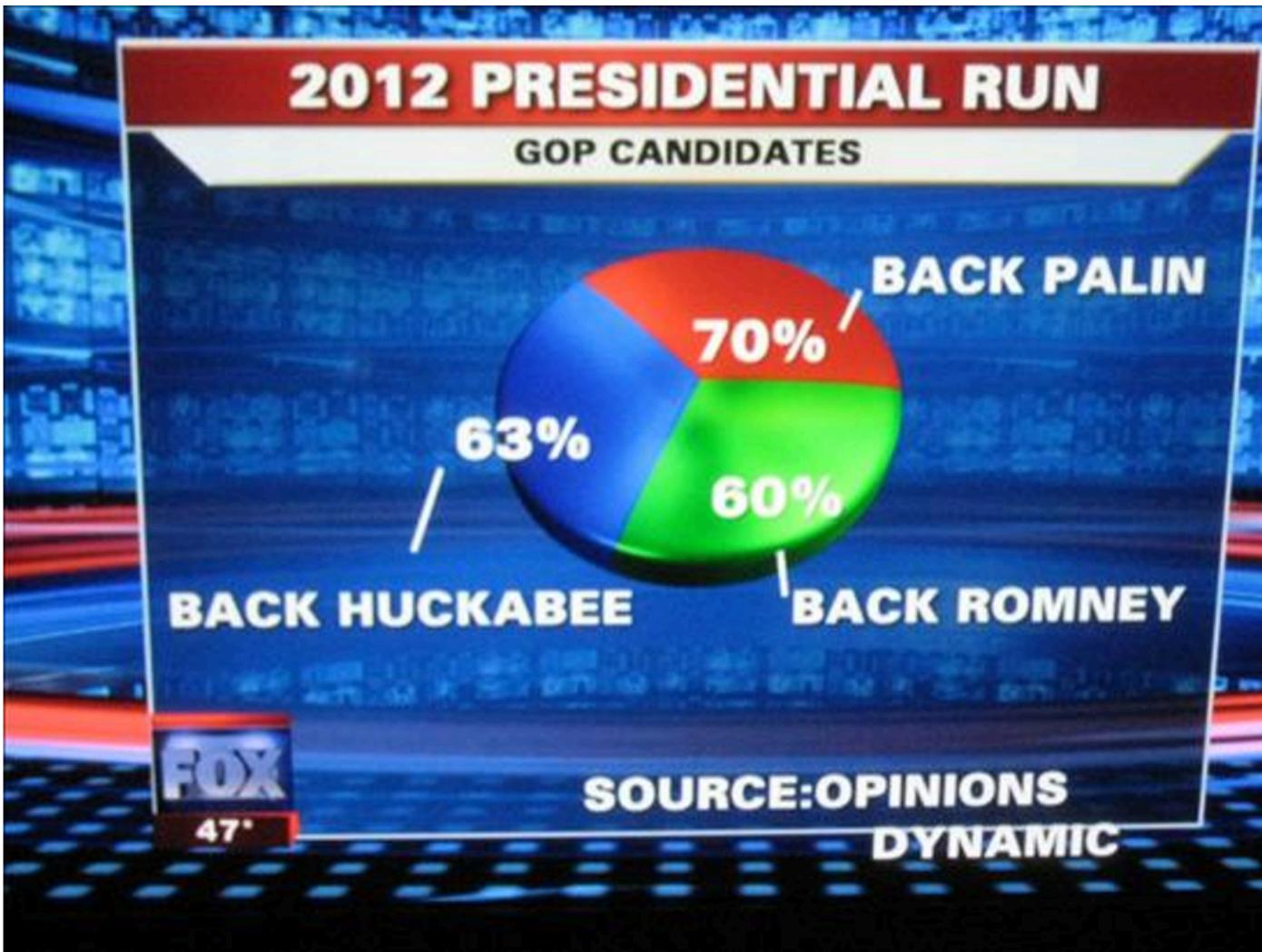
Total House and Senate campaign expenditures,
in millions



Source: Citizens' Research Foundation
TIME Chart by Nigel Holmes



(The same data presented with two different aspect ratios)



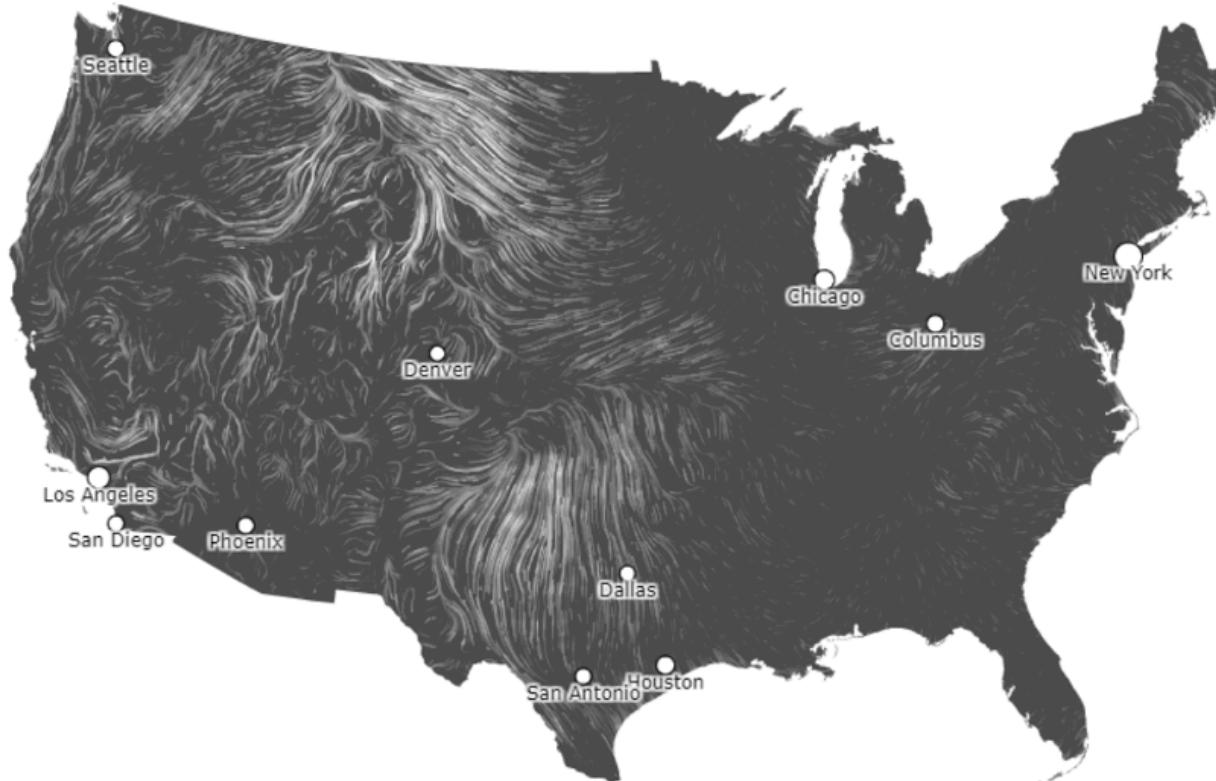
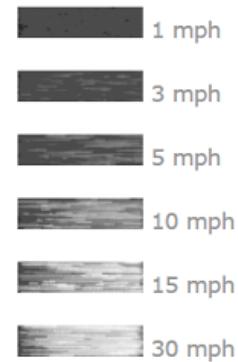
wind map

October 23, 2021

3:07 am EST

(time of forecast download)

top speed: 27.0 mph
average: 7.1 mph



(Click image to view interactive visualization)

Activity

- Each of the questions corresponds to an important quality of data visualizations:
 - Is the visualization pleasing to look at? → **Aesthetic**
 - Does the visualization accurately and honestly present data? → **Substantive**
 - Can we understand what message the maker of the visualization is attempting to convey? → **Perceptual**
- We need to consider all of these qualities when evaluating and designing 'good' data visualizations

What data visualization IS

- Dependent on:
 - **Context** → *where and how* will our visualization be used? (eg. academic journal, poster, infographic)
 - **Audience** → *who* is intended to use our visualization? (eg. subject experts, general public)
 - **Data structure** → *what* information do our data capture? (eg. quantities, relationships)

What data visualization is NOT

- Hard and fast rules for every situation → visualizing data means making decisions

What tools are used for data visualization?

Tools for Data Visualization

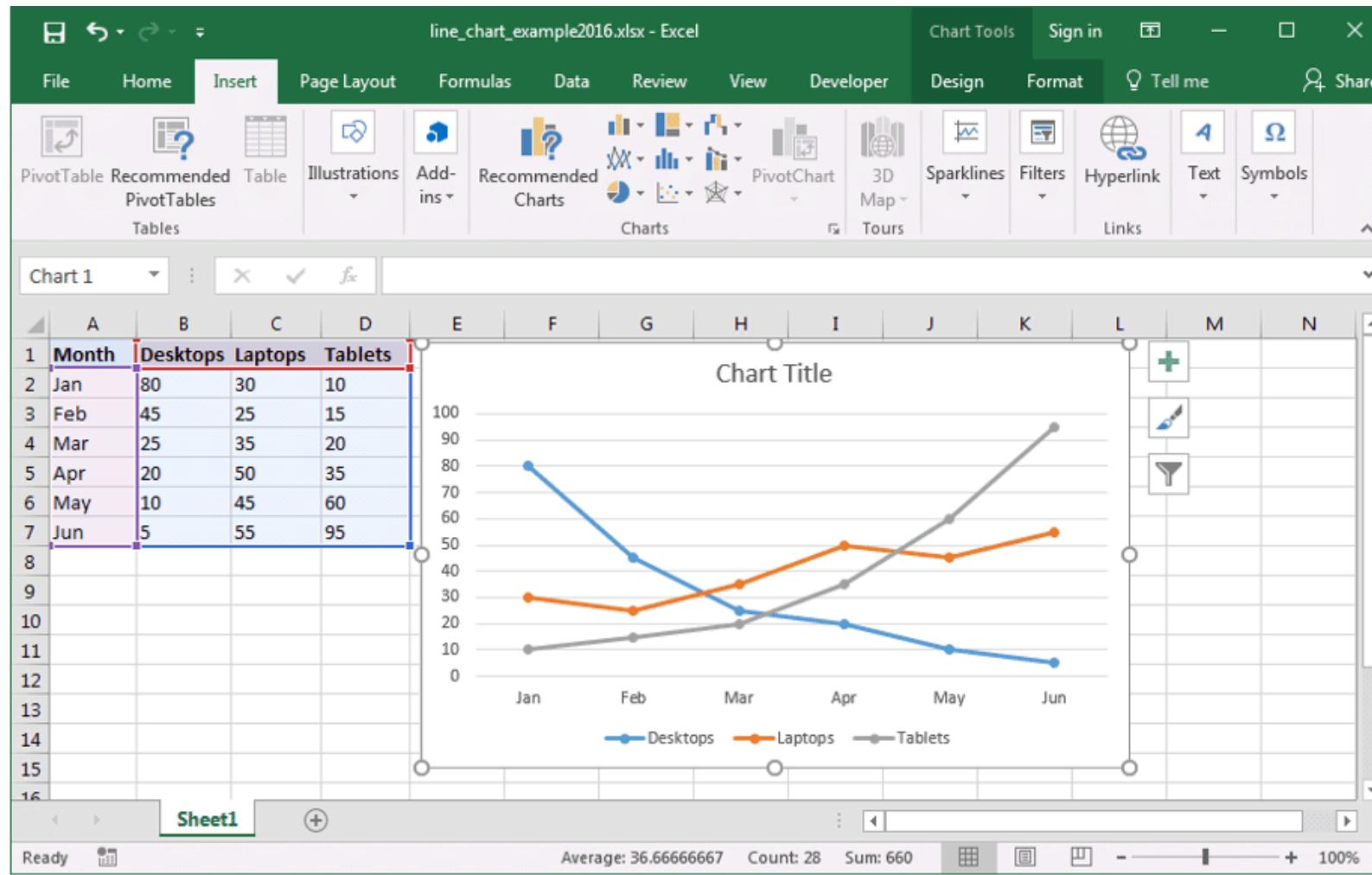


Microsoft Excel (LibreOffice Calc, Google Sheets, etc)



| | |
|-----------------------------|--|
| What is it | Spreadsheet software with ability to generate static data visualizations |
| Access | <ul style="list-style-type: none">- Excel is paid (part of MS Office Suite)- Free alternatives such as Google Sheets and LibreOffice Calc |
| Reproducible visualizations | <ul style="list-style-type: none">- no |
| Ease of use | <ul style="list-style-type: none">- Point and click to select from pre-made visualizations- Can serve as frontend for databases using Power Query |
| Use cases | <ul style="list-style-type: none">- "First line tool" for data analysis and visualization- pretty much everywhere |

Microsoft Excel (LibreOffice Calc, Google Sheets, etc)



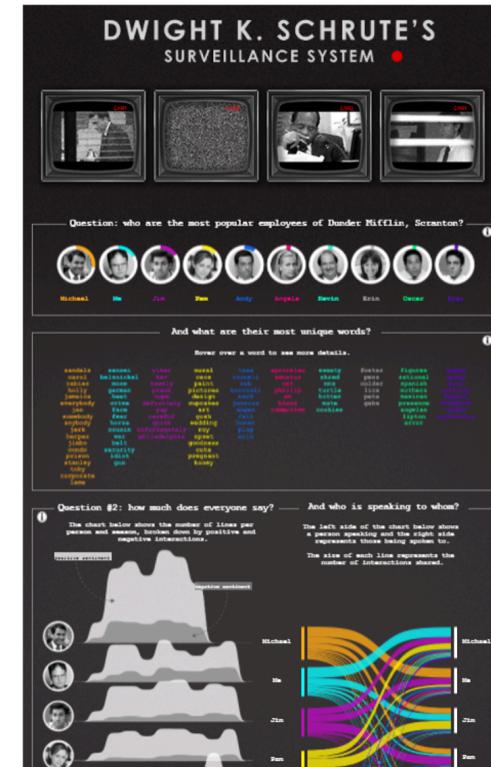
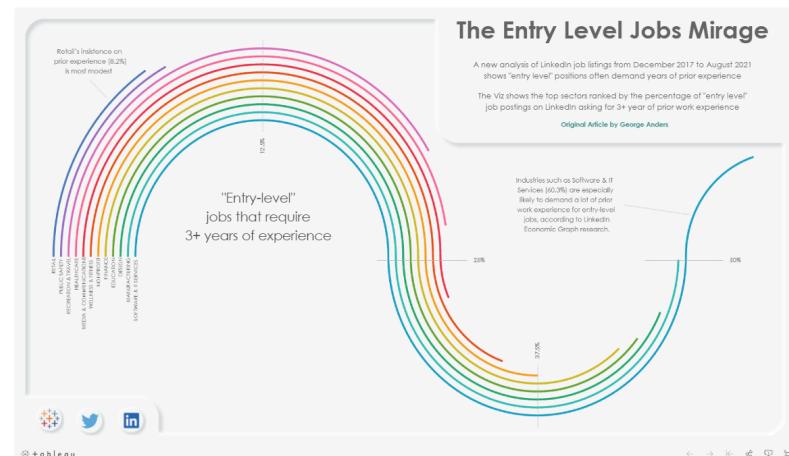
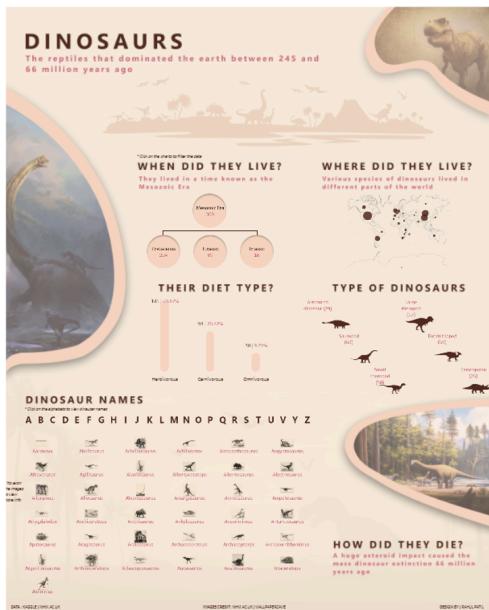
Tableau, Tableau Public



| | |
|-----------------------------|---|
| What is it | Combines data from different sources (databases, spreadsheets) into interactive, dynamic visualizations on the web |
| Access | <ul style="list-style-type: none">- Tableau Server and Desktop are paid- Tableau Public is free but all visualizations are public and not saved locally |
| Reproducible visualizations | <ul style="list-style-type: none">- no |
| Ease of use | <ul style="list-style-type: none">- Point and click to select from pre-made visualizations |
| Use cases | <ul style="list-style-type: none">- Industry (designed for business intelligence), infographics for media |

Tableau, Tableau Public

Click each link to view and interact with public visualizations chosen as Tableau's 'Viz of the Day':
🔗 left, 🔗 middle, 🔗 right





Microsoft Power BI

| | |
|-----------------------------|---|
| What is it | Combines data from different sources (databases, spreadsheets) into interactive, dynamic visualizations on the web |
| Access | Paid (part of MS Office Suite) |
| Reproducible visualizations | no |
| Ease of use | Drag and drop to select from pre-made visualizations Can use DAX (Data Analysis Expressions) functions to perform operations on data |
| Use cases | Industry, government (designed for business intelligence) |

Microsoft Power BI

Sales Report - Power BI Desktop

Kaur Kotadia

File Home Insert Modeling View Help

Cut Copy Format painter Paste

Formatting Segoe UI B I U A A

Get data Excel SQL Server data source Data Transform Refresh data New visual Text box More

Queries Insert New measure Quick Calculations Publish Share

OVERVIEW

Sales Report

\$5.3M Australia \$5.3M Canada \$2.6M France \$2.3M Germany \$3.3M UK \$21.8M USA

Key influencers Top segments

What influences NPS to be 7

When ...

- UnitPrice is 298.5 - 299.94 → 10.20x
- UnitPrice is 197.45 - 199.45 → 10.20x
- Manufacturer is Litware, INC. → 10.20x
- Color is Brown → 10.20x
- StockType is High → 10.20x
- Manufacturer is Contoso, Ltd → 10.20x
- Color is Silver → 10.20x

Units by Product and Sale Size

Sales Amount by Brand Name

Units

SalesAmount

2013 February Contoso Prosswear Adventure Works Other Wide World Importers Northwind Th...

2014 Qtr 1 2014 Qtr 2 2014 Qtr 3 2014 Qtr 4

Visualizations Fields

Values Drag data fields here

Drillthrough Keep all filters Off Drag data fields here

PAGE 1 of 1

The screenshot displays the Microsoft Power BI Desktop application window. The top navigation bar includes File, Home, Insert, Modeling, View, and Help. The Home tab is selected. The ribbon below the menu has sections for Paste, Cut, Copy, Format painter, Get data (with options for Excel, SQL, and Server data source), Transform, Refresh data, New visual, Text box, More, New measure, Quick Calculations, Publish, and Share. On the left, there's a navigation pane with icons for Overview, Stores, Products, and a plus sign. The main area contains several visualizations: 1) An 'OVERVIEW' section with six large numerical values representing sales for Australia, Canada, France, Germany, UK, and USA. 2) A 'Key influencers' section showing how various product attributes (UnitPrice, Manufacturer, Color, StockType) influence Net Promoter Score (NPS). 3) Three violin plots showing the distribution of units sold for Large, Medium, and Small products. 4) Two treemap charts showing Sales Amount by Brand Name, where Contoso, Fabrikam, and Litware are the largest segments. 5) Two stacked area charts showing Sales Amount by Year, Month, and Brand Name, with data points for Contoso, Prosswear, Adventure Works, and others. The right side of the interface features a 'Visualizations' pane with a grid of icons for different chart types, and a 'Fields' pane where users can drag data fields to define values or drillthrough. The bottom status bar indicates 'PAGE 1 of 1'.

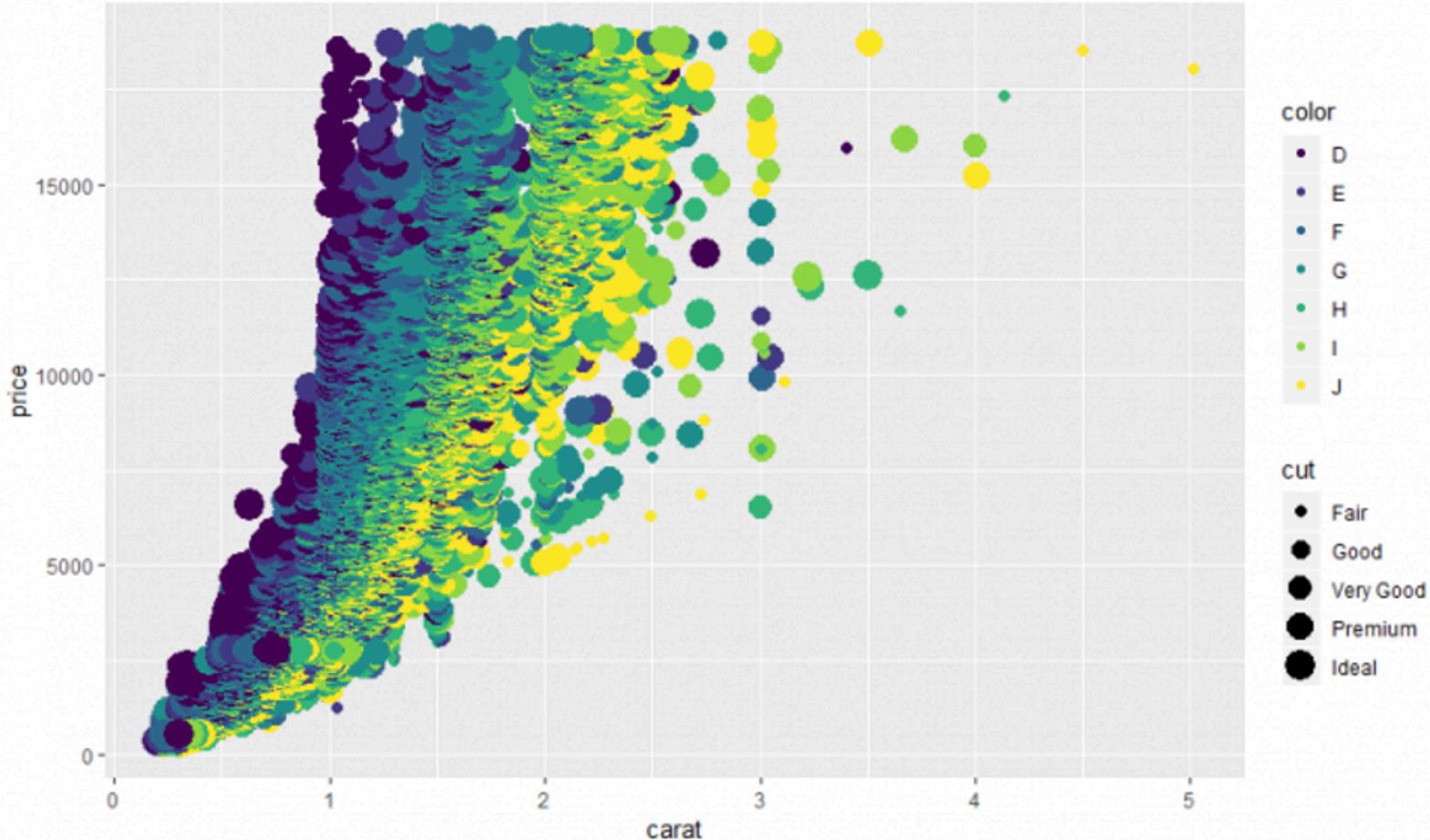


R

| | |
|-----------------------------|--|
| What is it | Programming language with libraries for data visualization (e.g., ggplot2, Plotly, RColorBrewer) |
| Access | Free and open source (https://www.r-project.org/COPYING) |
| Reproducible visualizations | yes |
| Ease of use | Programming language; requires some coding knowledge |
| Use cases | Academia, industry, government; research and data science contexts |

R

Scatterplot example on diamonds dataset

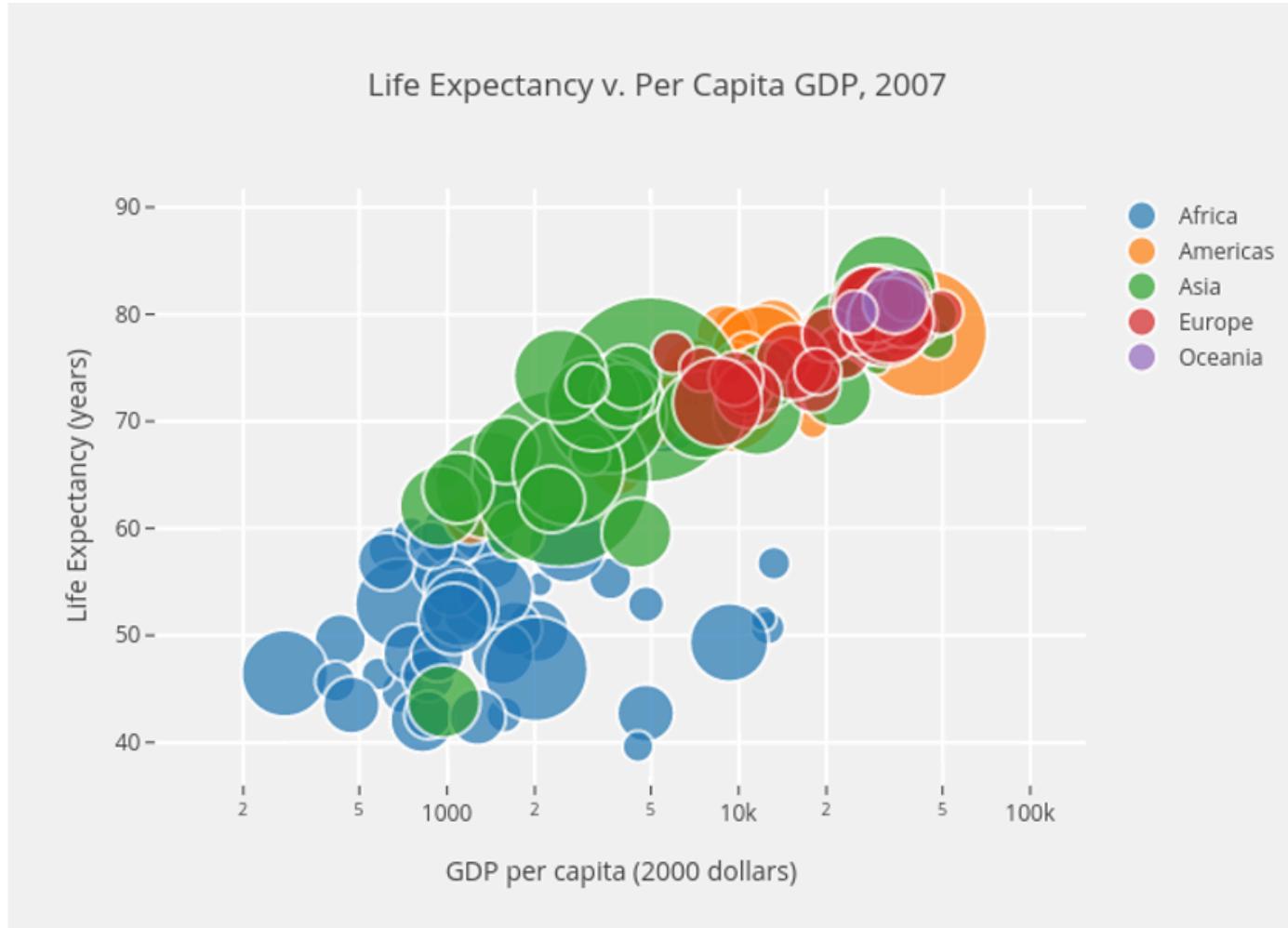


Python



| | |
|-----------------------------|---|
| What is it | Programming language with libraries for data visualization (e.g., Matplotlib, Plotly) |
| Access | Free and open source Python Software Foundation License |
| Reproducible visualizations | yes |
| Ease of use | Programming language; requires some coding knowledge |
| Use cases | Government, industry, academia; data science and programming contexts |

Python



For our purposes

- Step-by-step walkthroughs are focused on Python (Matplotlib)
 - [Commonly used tool](#)
 - Free and open source
 - Reproducible
 - LOTS of available resources online

BUT...

- General design principles will apply to creating data visualizations in whichever software you decide to use

Some admin

- Github overview
- Schedule & What will we cover in this course?
- Assignment 1
- LLMS vs Stack Overflow

Coming Up in this Course - Topics

- First Steps

- Get started
- Make a plot (Matplotlib)
- Thinking about reproducibility

- Graphing Our Data

- Customize our plot appearance
- Choosing the right visualization (perceptual qualities)

Coming Up in this Course - Topics

- **Visualization with Purpose**
 - Subplots and combining visualizations
 - Colour theory and accessible design
 - Data visualization as advocacy
- **Getting Fancy**
 - Beyond matplotlib (Seaborn)
 - Drawing maps (geoplot)
 - Interactive data visualizations (Plotly)

Learning Outcomes of this Course

1. Create and customize data visualizations start to finish in Python
2. Use general design principles for creating accessible and equitable data visualizations in Python and other software
3. Use data visualization to purposely tell a story