Data Visualization

Introduction and Overview

Ciara Zogheib

Data Sciences Institute, University of Toronto

I 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 the Mississaugas of the Credit. 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.

Hello!

- Ciara Zogheib (She/Her)
- ciara.zogheib@mail.utoronto.ca
- PhD Student at the Faculty of Information
- Working as a data scientist/researcher with the government
- Started doing data science by accident because I wanted to go to Greece

Introduce yourself in the chat!

Prerequisites

- R and RStudio are installed on your computer
- **Question:** Who has experience using R?

Assignments

Assignment 1: Participation (Attendance + Class Content)	10%
Assignment 2: Good and Bad Data Visualization	30%
Assignment 3: Data Visualization Ethics	20%
Assignment 4: Final Project	40%

Submitting your work

- Create a folder in Google Drive or GitHub (whichever you prefer)
  - Name it DSIDataViz_YourFirstNameYourLastName
- Send the link to the folder to ciara.zogheib@mail.utoronto.ca
- **Please** make sure assignments and notes are clearly labelled!

About grading and tutorials...

Meet your TA

Today, we will...

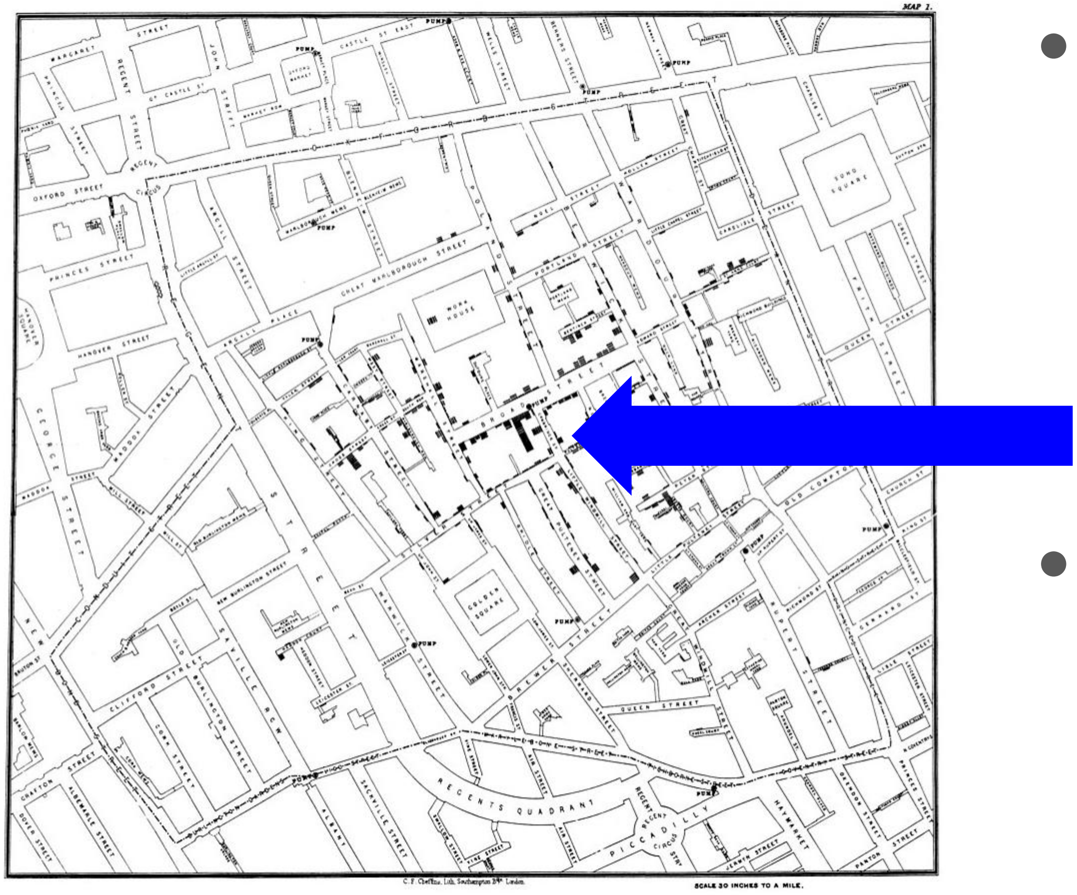
- Explore why we should care about data visualization
- Question what makes ‘good’ data visualization
- Introduce a range of software and tools that are used for data visualization
- Go through an overview of what to expect in the rest of this course

Case Study: Why should we care about data visualization?

- 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

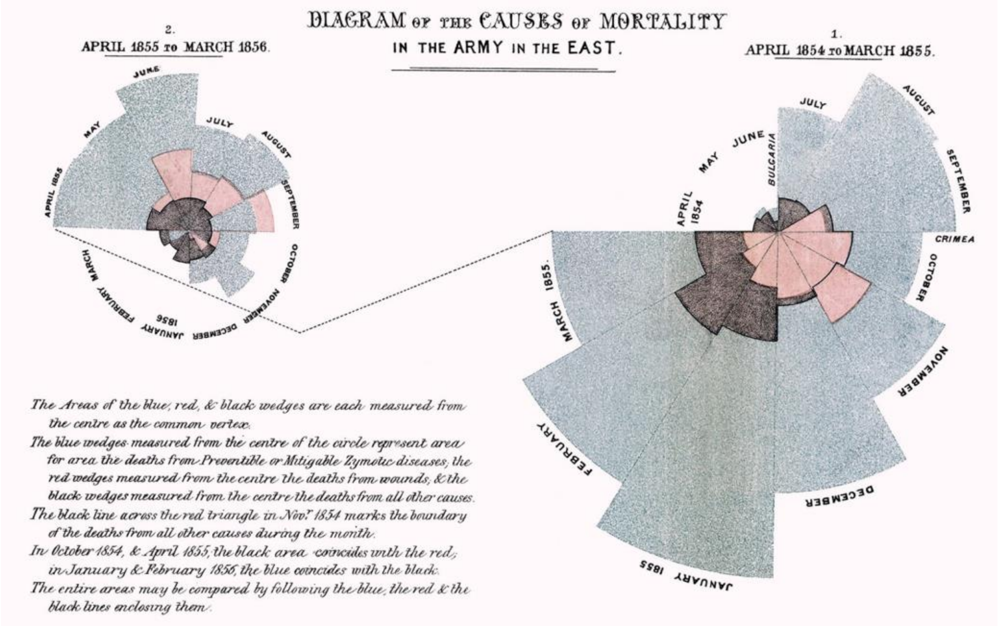


DIAGRAM OF THE CAUSES OF MORTALITY

IN THE ARMY IN THE EAST.

The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases; the red wedges measured from the centre the deaths from wounds; & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Novr. 1854 marks the boundary of the deaths from all other causes during the month.

In October 1854, & April 1855, the black area coincides with the red; in January & February 1856, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.

- 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

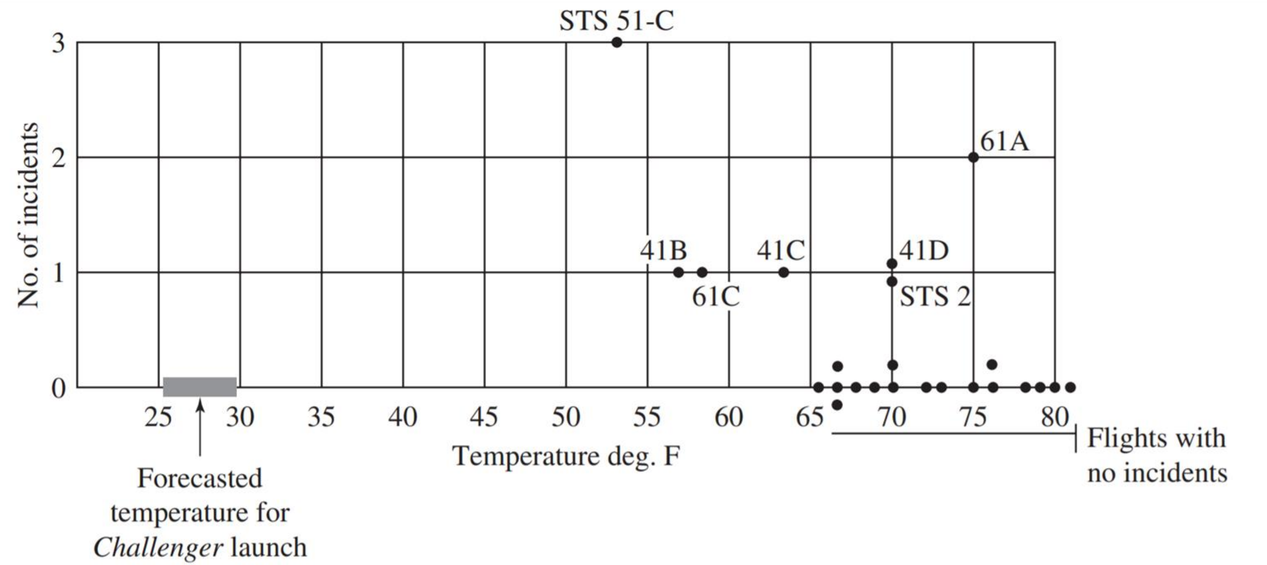
But data visualization is not always so straightforward…

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

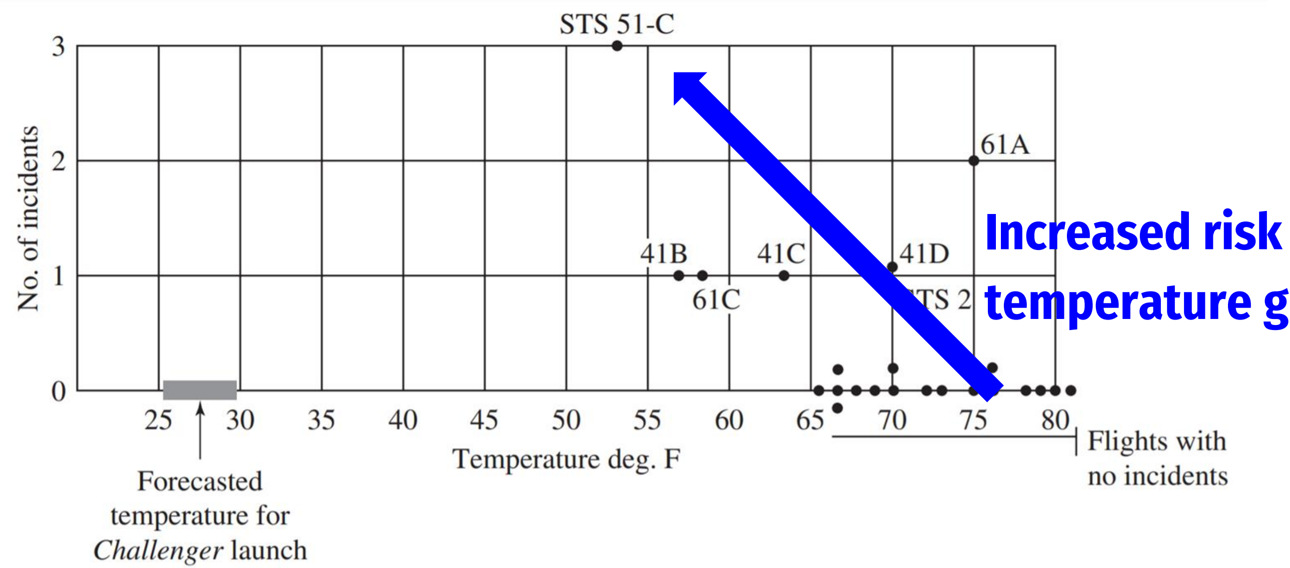
- In 1997, Edward Tufte famously published a visualization showing the relationship between temperature during launches of test shuttle flights (pre-*Challenger*) and incidents of damage to O-rings



(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*



Increased risk to O-rings as temperature goes down

(Tufte, 1997)

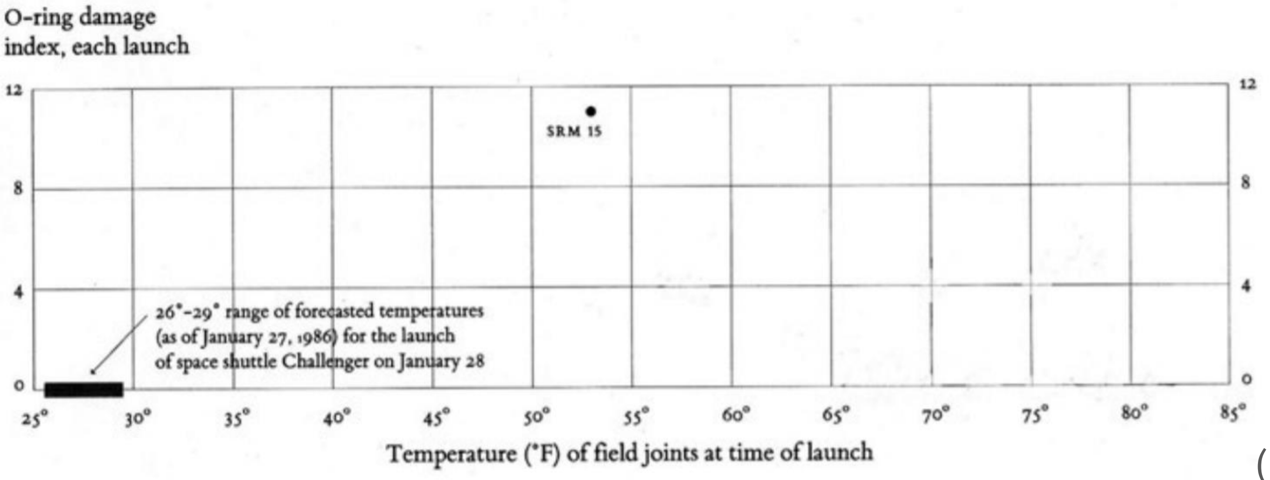
The Challenger Disaster - Tufte’s Graph

- 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?

(Tufte, 1997)

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:



26°-29° range of forecasted temperatures (as of January 27, 1986) for the launch of space shuttle Challenger on January 28

(Robison, 1997, 2002)

The Challenger Disaster - But…

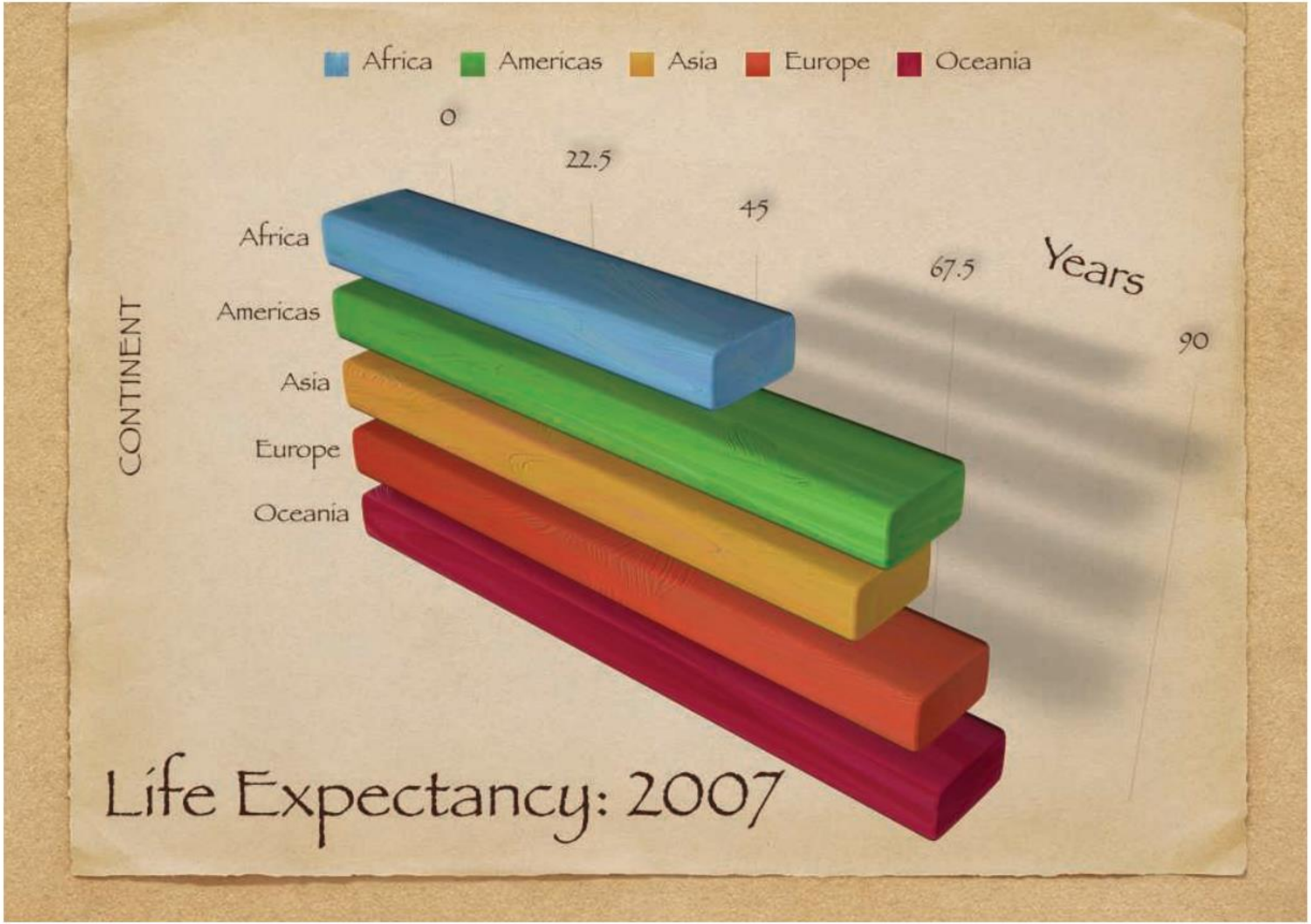
- Per Tufte, the *Challenger* disaster is a case where good data visualization could have facilitated understanding of complex data and saved lives
- Per Robison, Tufte's work is a case of bad data visualization unfairly placing blame and leading the audience to a faulty conclusion

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

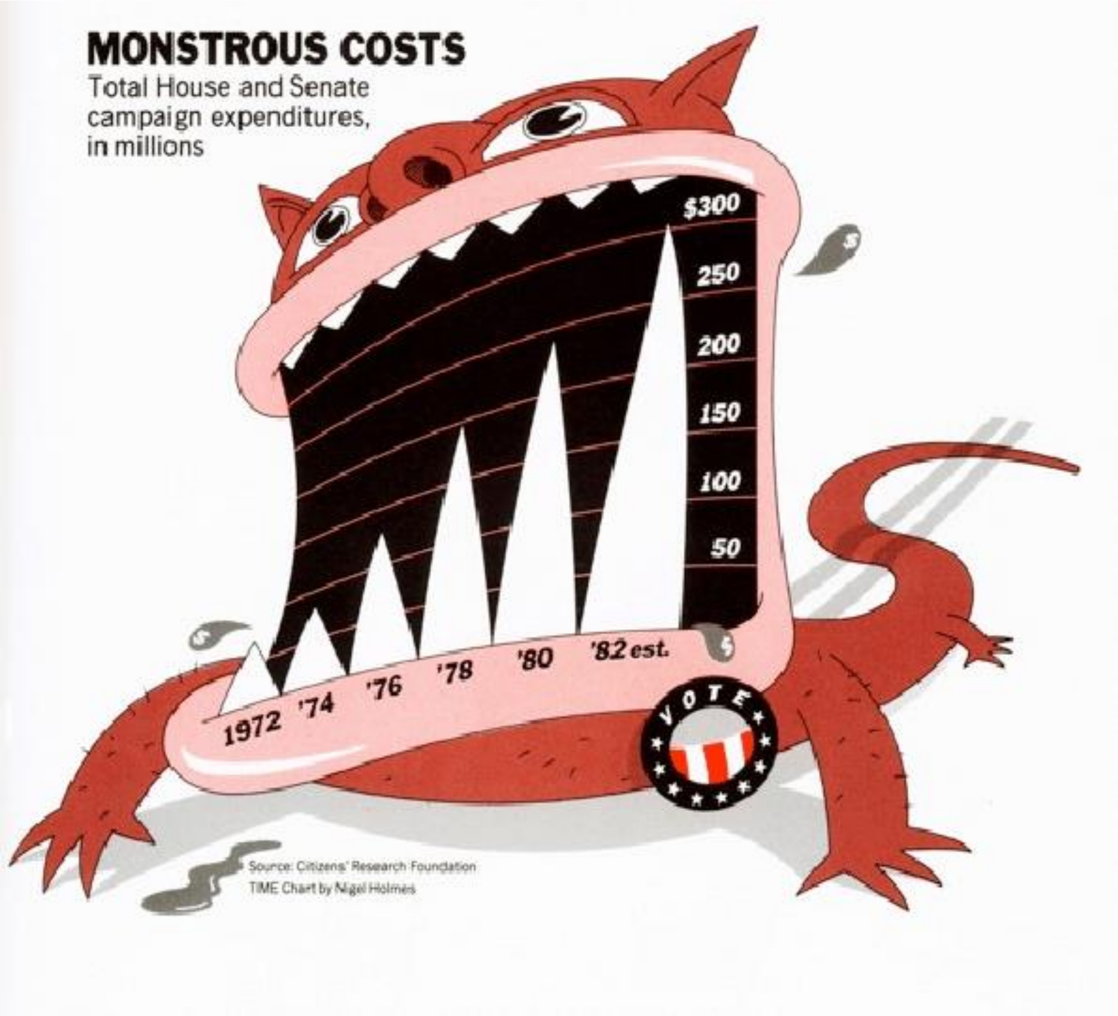
Activity: What is ‘good’ data visualization?

Activity

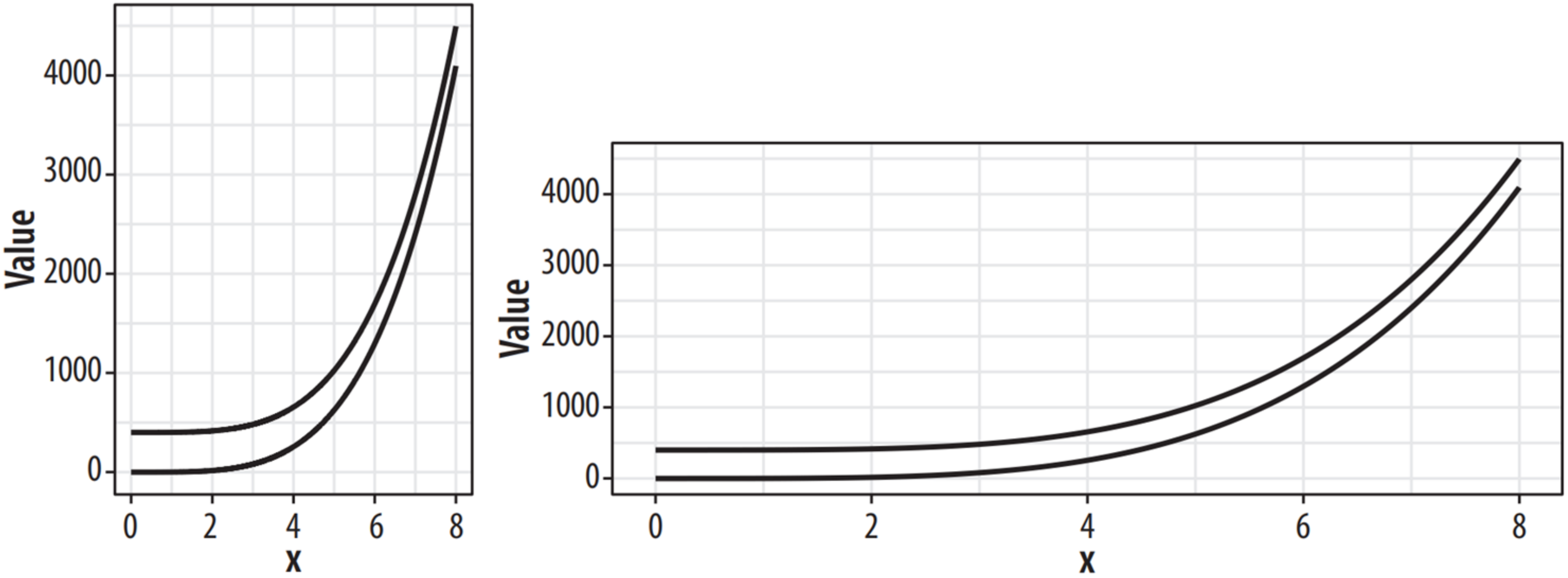
- Look at the following examples of data visualizations. For each, consider:
  - Is the visualization pleasing to look at?
  - Does the visualization accurately represent data?
  - Can we understand what message the maker of the visualization is attempting to convey?
  - Consider factors such as colour,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


Africa
Americas
Asia
Europe
Oceania
0
22.5
45
67.5
90
Years
CONTINENT
Africa
Americas
Asia
Europe
Oceania
Life Expectancy: 2007

(Healy, 2018)

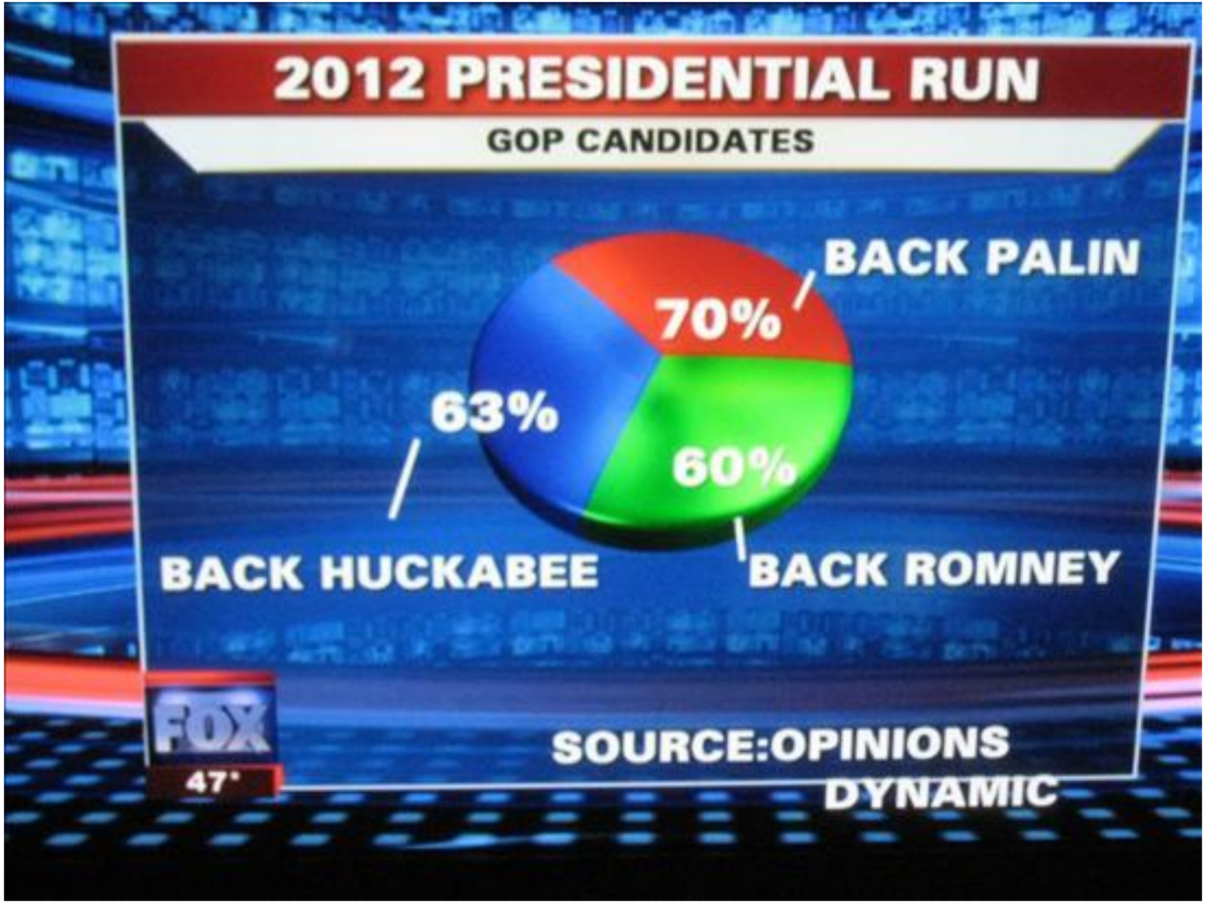

MONSTROUS COSTS
Total House and Senate campaign expenditures, in millions
$300
250
200
150
100
50
1972
'74
'76
'78
'80
'82 est.
VOTE
Source: Citizens' Research Foundation
TIME Chart by Nigel Holmes

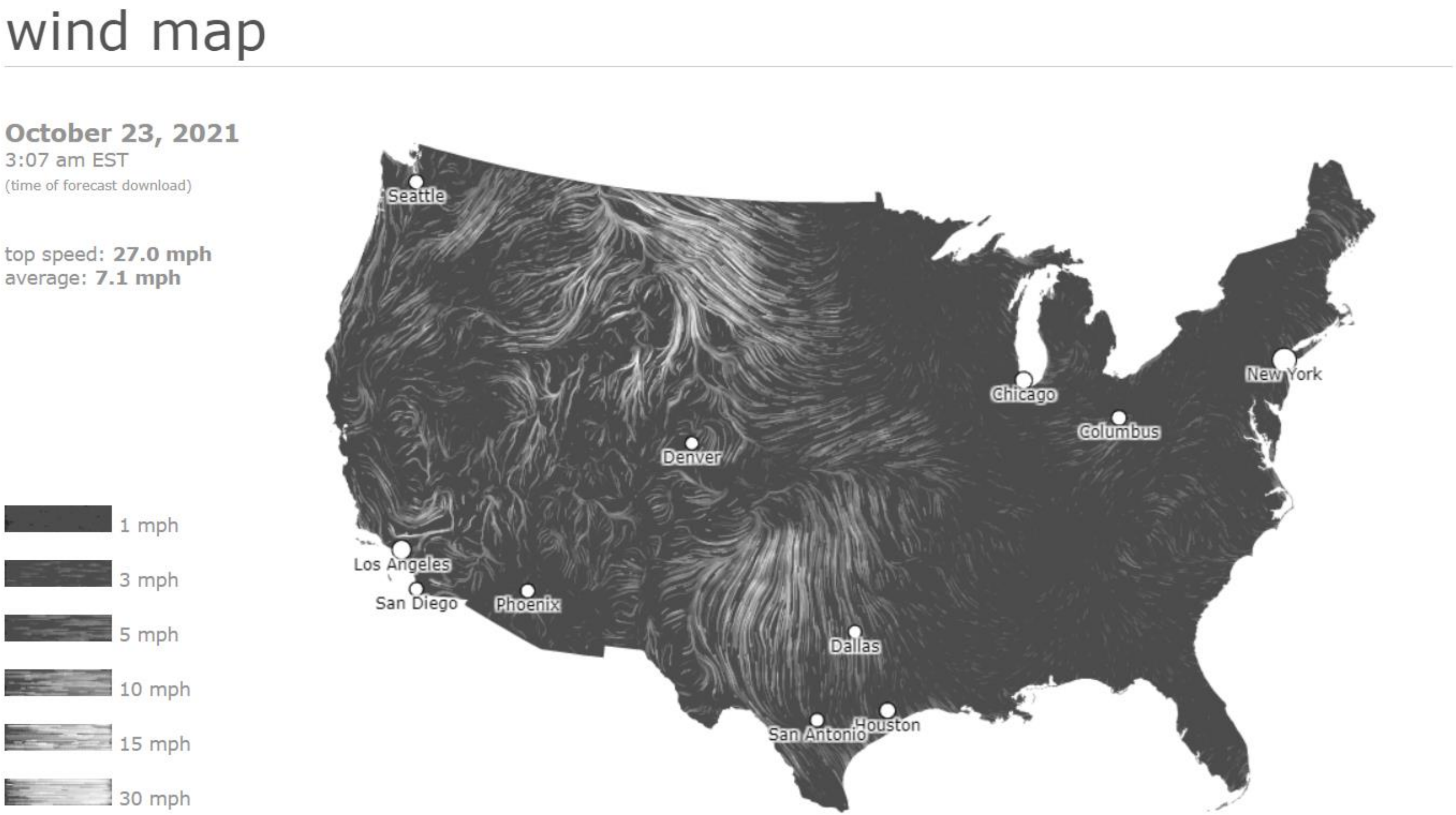
(Healy, 2018)



(Healy, 2018)

(The same data presented with two different aspect ratios)


2012 PRESIDENTIAL RUN
GOP CANDIDATES
BACK PALIN
70%
63%
60%
BACK HUCKABEE
BACK ROMNEY
FOX
47°
SOURCE:OPINIONS
DYNAMIC



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

(Healy, 2018; Chapter 1)

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)

(Healy, 2018; Chapter 1)

What data visualization is NOT

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

(Healy, 2018; Chapter 1)

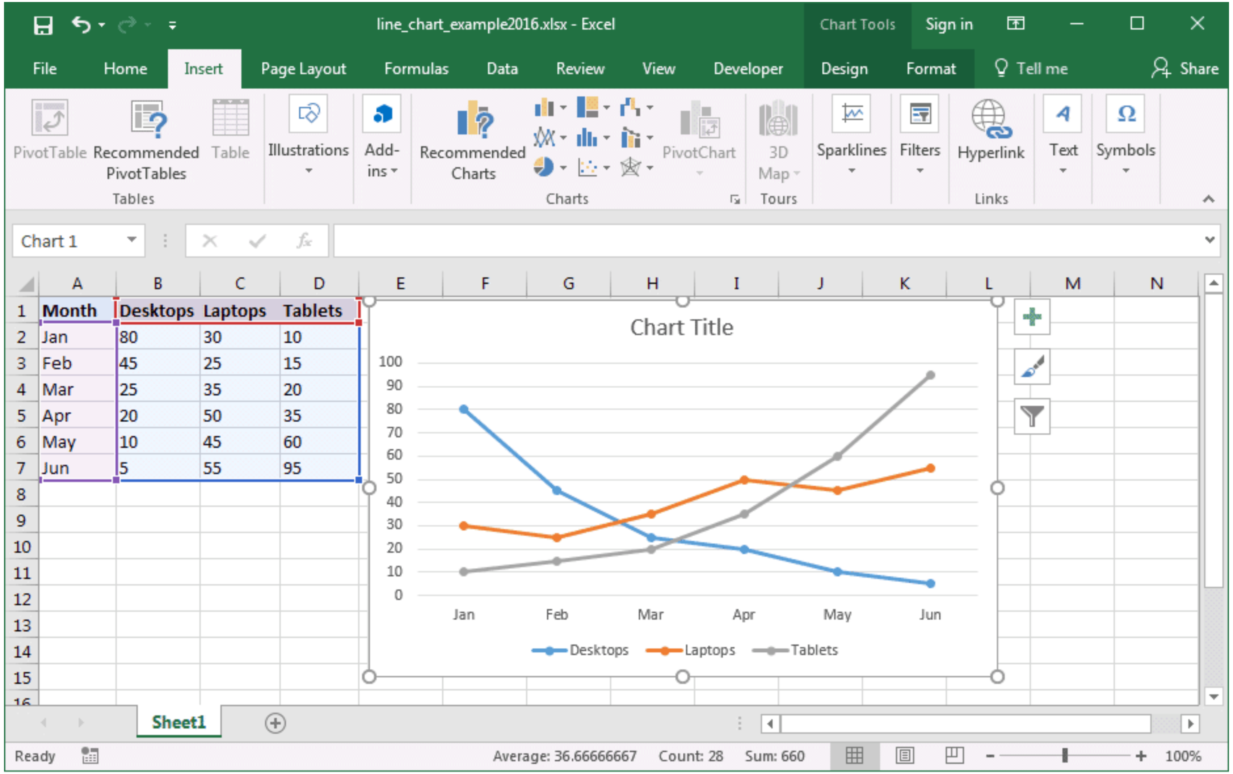
What tools are used for data visualization?

Tools for Data Visualization

Microsoft Excel (LibreOffice Calc, Google Sheets, etc)

What is it:	● Spreadsheet software with ability to generate static data visualizations
Access:	● Excel is paid (part of MS Office Suite) ● Free alternatives such as Google Sheets and LibreOffice Calc
Reproducible visualizations:	● No
Ease of use:	● Point and click to select from pre-made visualizations ● Can serve as frontend for databases using Power Query
Use cases:	● “First line tool” for data analysis and visualization - pretty much everywhere

Microsoft Excel (LibreOffice Calc, Google Sheets, etc)



Month	Desktops	Laptops	Tablets
Jan	80	30	10
Feb	45	25	15
Mar	25	35	20
Apr	20	50	35
May	10	45	60
Jun	5	55	95

Average: 36.66666667 Count: 28 Sum: 660

TechOnTheNet.com

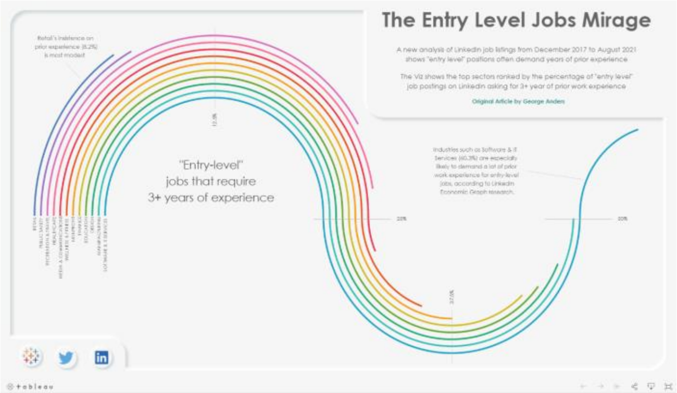
Tableau, Tableau Public

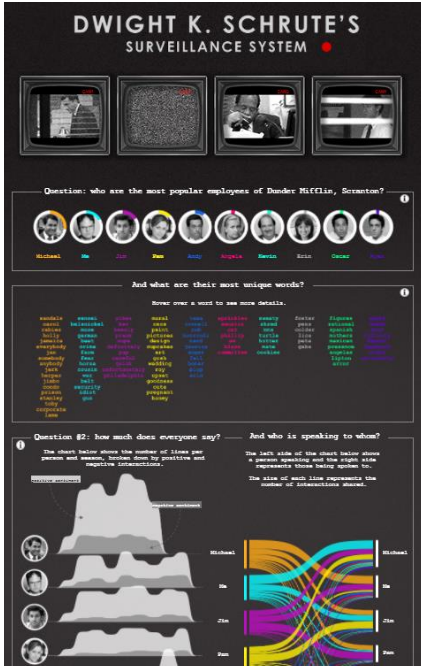
What is it:	● Combines data from different sources (databases, spreadsheets) into interactive, dynamic visualizations on the web
Access:	● Tableau Server and Desktop are paid ● Tableau Public is free **BUT** all visualizations are public and not saved locally
Reproducible visualizations:	● No
Ease of use:	● Point and click to select from pre-made visualizations
Use cases:	● Industry (designed for business intelligence), infographics for media

Tableau, Tableau Public

Click each image to view and interact with public visualizations chosen as Tableau's 'Viz of the Day':

DINOSAURS
The reptiles that dominated the earth between 245 and 66 million years ago
WHEN DID THEY LIVE?
WHERE DID THEY LIVE?
THEIR DIET TYPE?
TYPE OF DINOSAURS
DINOSAUR NAMES
A B C D E F G H I J K L M N O P Q R S T U V Y Z
HOW DID THEY DIE?
A huge asteroid impact caused the mass dinosaur extinction 66 million years ago


The Entry Level Jobs Mirage
A new analysis of LinkedIn job listings from December 2017 to August 2021 shows "entry level" positions often demand years of prior experience
The Viz shows the top sectors ranked by the percentage of "entry level" job postings on LinkedIn asking for 3+ year of prior work experience
Original Article by George Anders
"Entry-level" jobs that require 3+ years of experience
+ a b l e a u

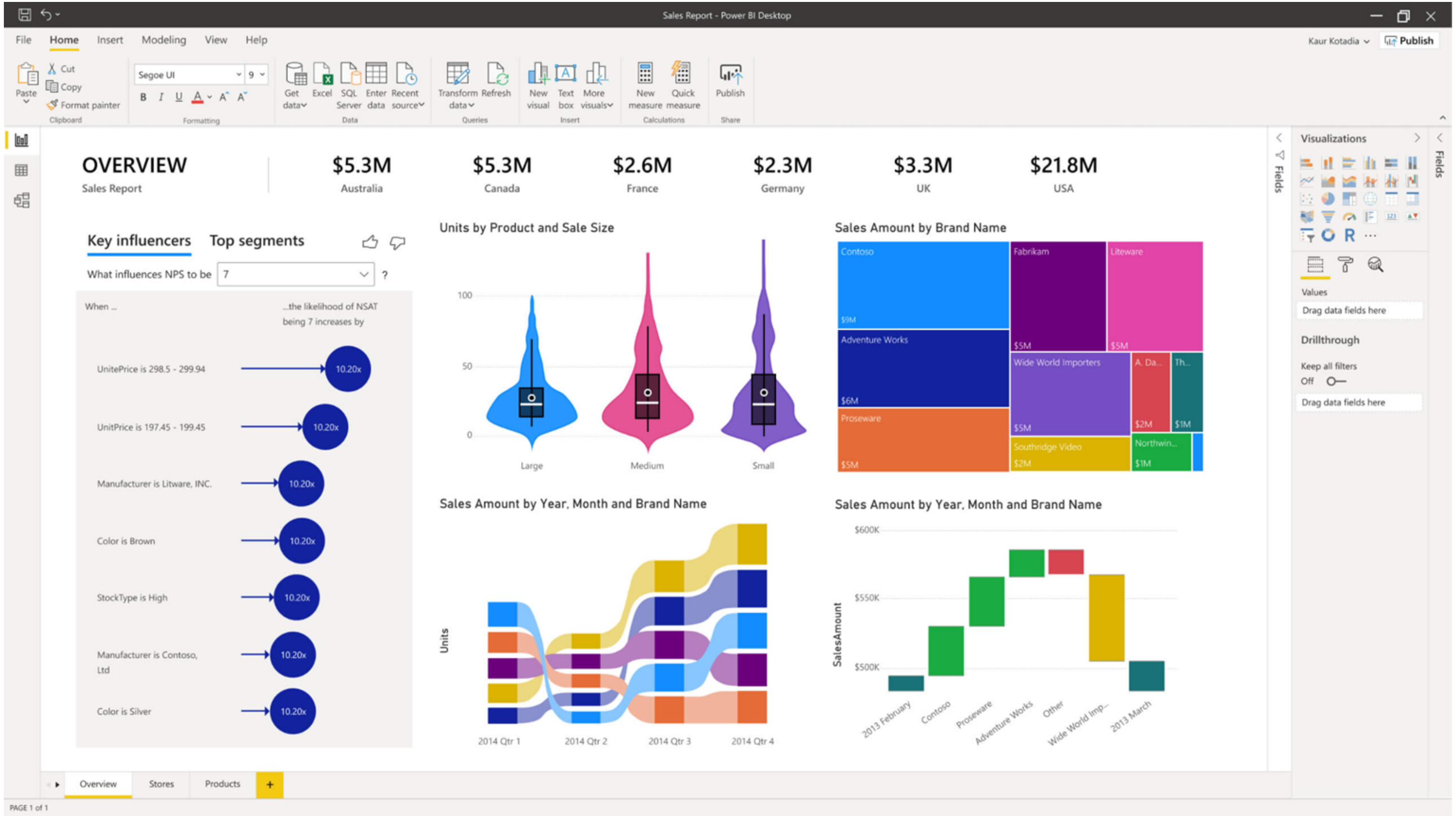

DWIGHT K. SCHRUTE'S
SURVEILLANCE SYSTEM
Question: who are the most popular employees of Dunder Mifflin, Scranton?
Michael
Jim
Pam
Andy
Angela
Kevin
Erin
Oscar
And what are their most unique words?
Hover over a word to see more details.
Question #2: how much does everyone say?
And who is speaking to whom?
Michael
Jim
Pam

Microsoft Power BI

Power BI

What is it:	● Combines data from different sources (databases, spreadsheets) into interactive, dynamic visualizations
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)

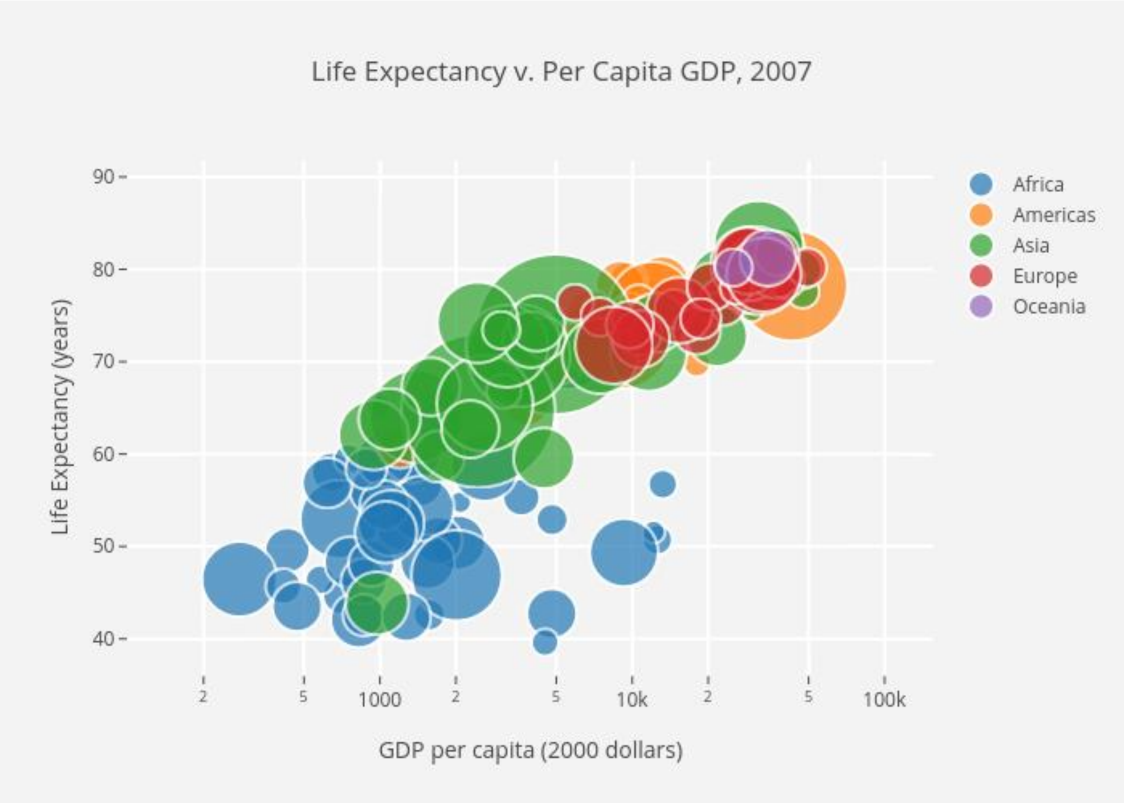
Microsoft Power BI



Python

What is it:	- Programming language with libraries for data visualization (eg. Matplotlib, Plotly)
Access:	- Free and open source (https://opensource.org/licenses/Python-2.0)
Reproducible visualizations:	- Yes
Ease of use:	- Programming language; requires some coding knowledge
Use cases:	- Government, industry, academia; data science and programming contexts

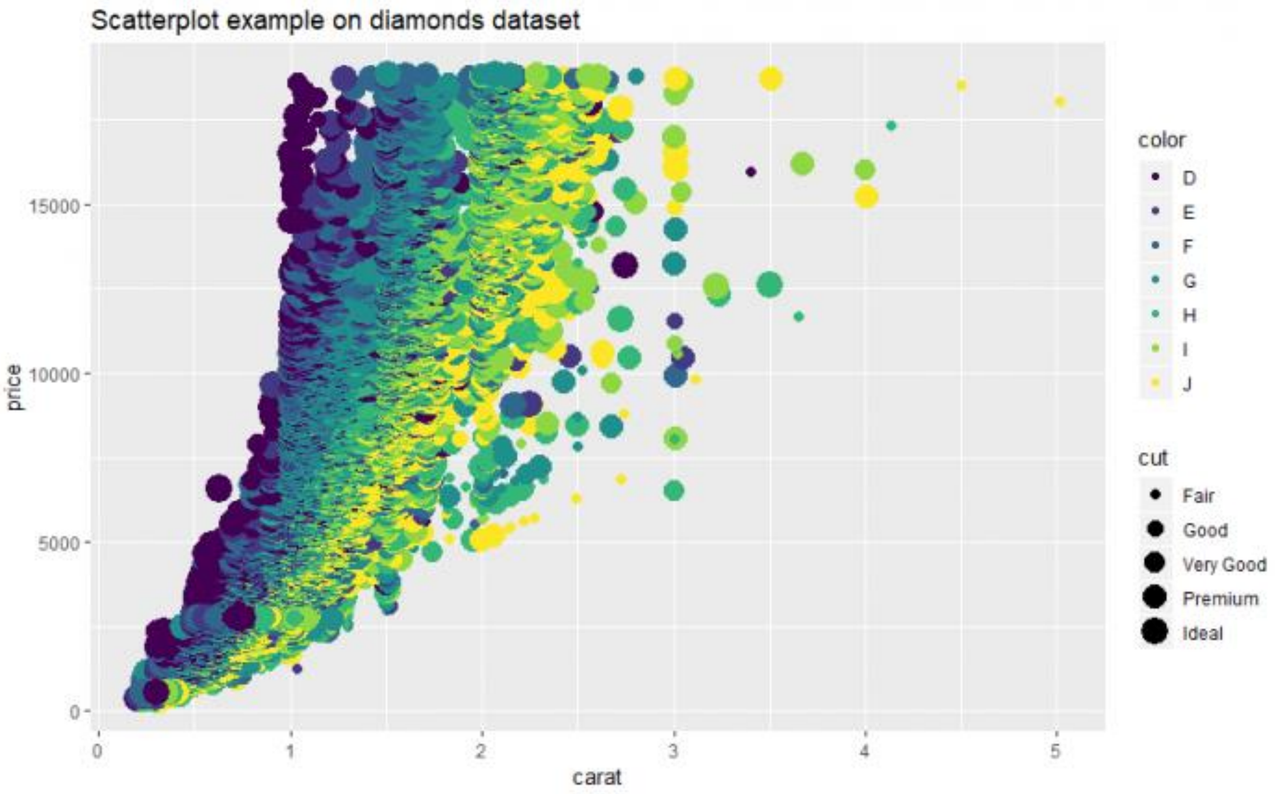
Python


Life Expectancy v. Per Capita GDP, 2007
Life Expectancy (years)
90
80
70
60
50
40
2
5
1000
2
5
10k
2
5
100k
GDP per capita (2000 dollars)
Africa
Americas
Asia
Europe
Oceania

R

What is it:	● Programming language with libraries for data visualization (eg. 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
price
0
5000
10000
15000
carat
0
1
2
3
4
5
color
D
E
F
G
H
I
J
cut
Fair
Good
Very Good
Premium
Ideal

For our purposes

- Step-by-step walkthrough and sample code are focused on R
  - Commonly used tool
  - Free and open source
  - Reproducible
  - LOTS of available resources online

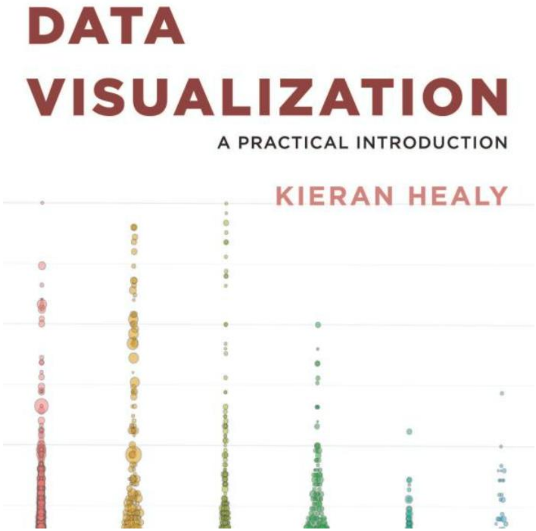
BUT...

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

What will we cover in this course?

Coming Up in this Course - Textbook

We will be following sample code from Healy's *Data Visualization: A Practical Introduction* (2018)



Coming Up in this Course - Topics

- **First Steps**
  - Get started
  - Make a plot (substantive qualities)
  - Thinking about reproducibility
- **Graphing Our Data**
  - Show the right numbers
  - Graph tables, add labels, make notes
  - Choosing the right visualization (perceptual qualities)

Coming Up in this Course - Topics

- **Visualization with Purpose**
  - Refine our plots (aesthetic qualities)
  - Colour theory and accessible design
  - Data visualization as advocacy
- **Getting Fancy**
  - Working with models; reproducibility
  - Drawing maps
  - Interactive data visualizations

Learning Objectives of this Course

1. Develop ability to **create and customize data visualizations** start to finish in R
2. Build an understanding of general design principles for creating **accessible and equitable** data visualizations in R and other software
3. Build an understanding of **data visualization as purposeful/telling a story** (and the ethical/professional implications thereof)