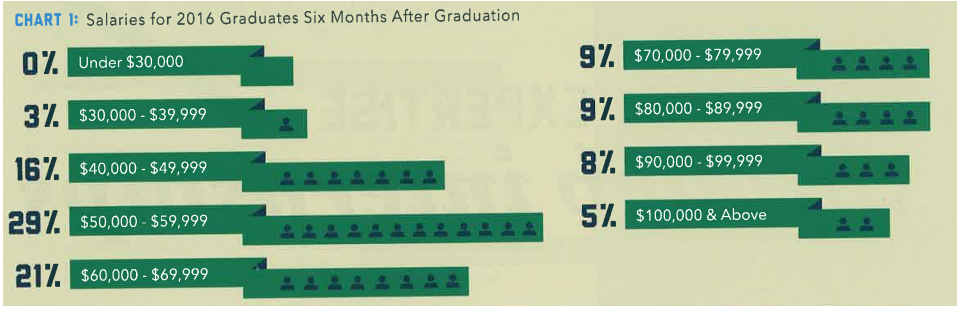
Lecture 1:
Welcome to Data Visualization Using R

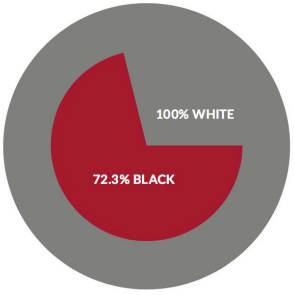
January 13, 2025

Take This Class So You Won't Make This Graphic



From Trachtenberg's 2018 magazine.

Or This One, 2 of 3

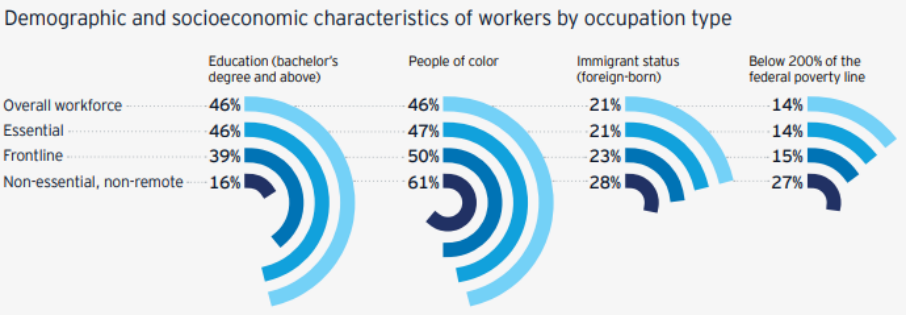


EQUALITY INDEX OF BLACK AMERICA, 2016–2017

	REVISED 2016	2017
EQUALITY INDEX	72.2%	72.3%
Economics	56.2%	56.5%
Health	79.4%	80.0%
Education	77.4%	78.2%
Social Justice	60.9%	57.4%
Civic Engagement	100.6%	100.6%

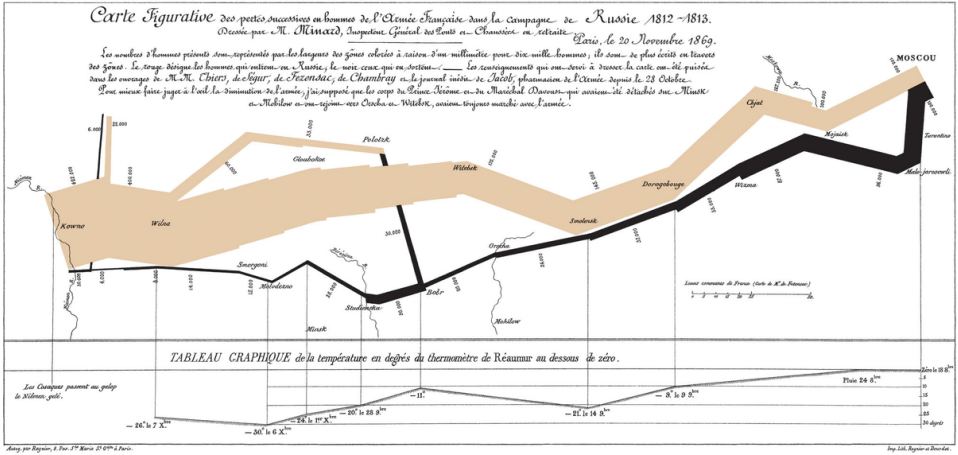
"U.S. Metros Ranked on Black-White Income Inequality," *Next City*, May 2, 2017

Or This One, 3 of 3



Remote Work in the Capital Region, 2021, Greater Washington Partnership.

Instead, Aspire to This



See Tufte for citation.

To Create Memories

- Journalists frequently start articles with anecdotes because they are
  - relateable
  - memorable
  - compelling (?)

To Create Memories

- Journalists frequently start articles with anecdotes because they are
  - relateable
  - memorable
  - compelling (?)
- Raw data is none of these things
- Goal of this course is to create graphics that are
  - compelling
  - clear
  - memorable
  - succinct

Course Administration

1. Write your name tent! I'll bring each week

Course Administration

1. Write your name tent! I'll bring each week
2. Syllabus
   - Policy brief handout
   - Fully composed chart handout
   - Good/bad/ugly assignments handout

Course Administration

1. Write your name tent! I'll bring each week
2. Syllabus
   - Policy brief handout
   - Fully composed chart handout
   - Good/bad/ugly assignments handout
3. Questions/issues with readings?

Course Administration

1. Write your name tent! I'll bring each week
2. Syllabus
   - Policy brief handout
   - Fully composed chart handout
   - Good/bad/ugly assignments handout
3. Questions/issues with readings?
4. Sign up for Piazza – email will go out today

Course Administration

1. Write your name tent! I'll bring each week
2. Syllabus
   - Policy brief handout
   - Fully composed chart handout
   - Good/bad/ugly assignments handout
3. Questions/issues with readings?
4. Sign up for Piazza – email will go out today
5. Introductions
   - name and degree
   - why this course?
   - what you do now
   - what you'd like to do when you're done

Today

1. R examples
2. Tufte
3. Getting started with R
4. R tools

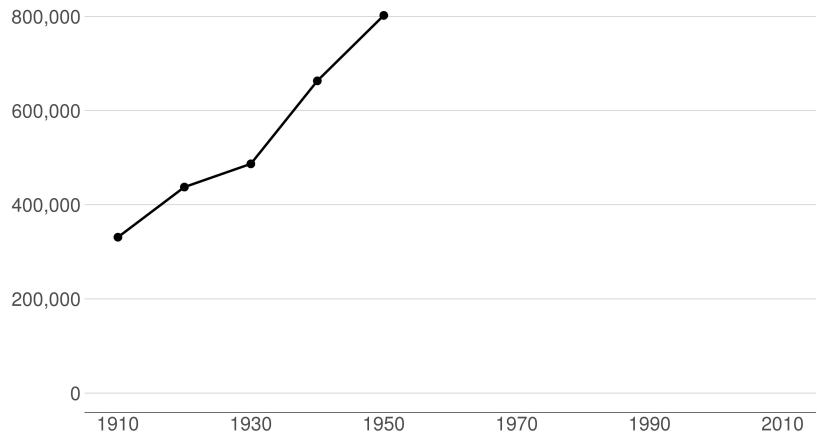
R Examples

R Examples

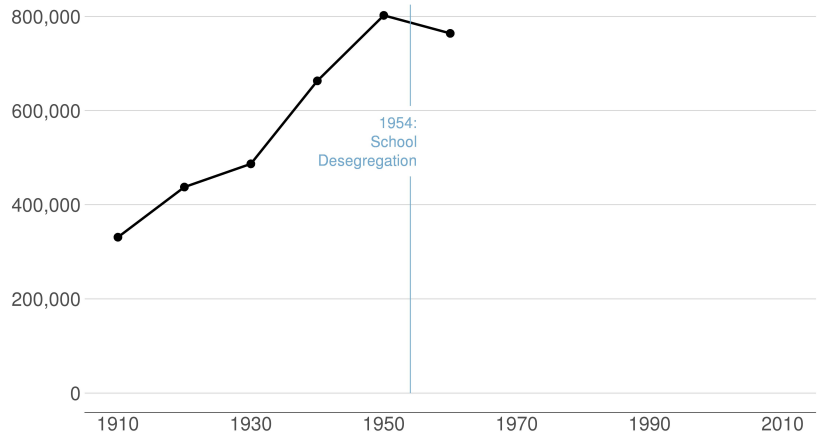
1. From a project about the long-run impacts of DC's 1968 civil disturbance
2. From a project about whether and why infrastructure costs are increasing
3. From a project about working from home in the DC region

From a Project about the Long-Run Impacts of DC's 1968 Civil Disturbance

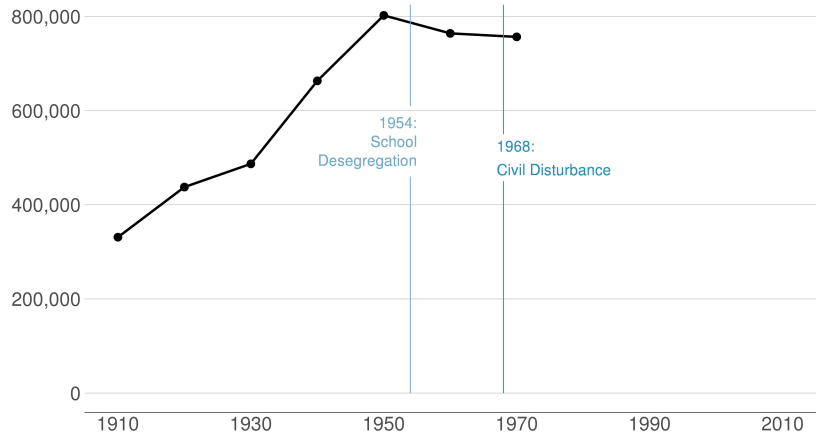
DC Gains Population Through 1950



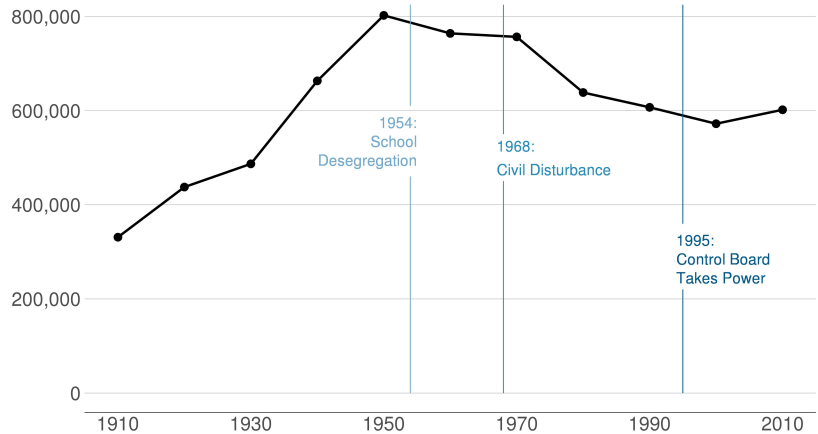
Population Loses Start with Desegregation



Continue After Civil Disturbance

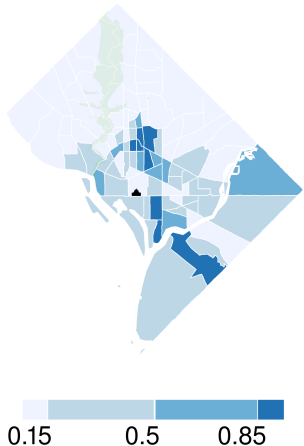


Population Turns Up After 2000



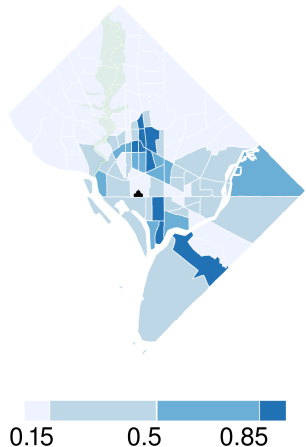
Profound Changes: Share African American by Neighborhood

1930
1940
1950
1960
1970
1980
1990
2000
2010



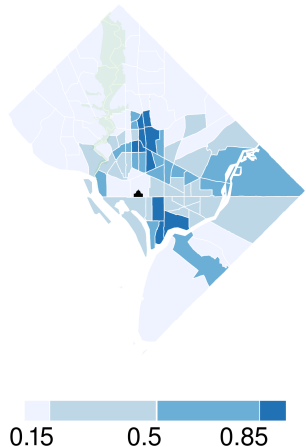
Profound Changes: Share African American by Neighborhood

1930
1940
1950
1960
1970
1980
1990
2000
2010



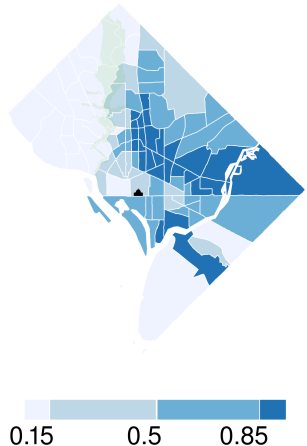
Profound Changes: Share African American by Neighborhood

1930
1940
1950
1960
1970
1980
1990
2000
2010



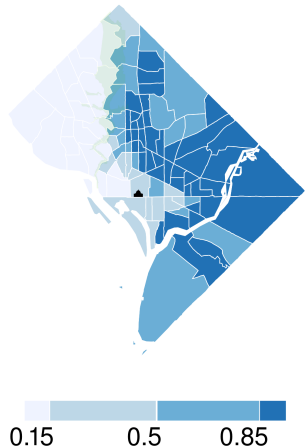
Profound Changes: Share African American by Neighborhood

1930
1940
1950
1960
1970
1980
1990
2000
2010



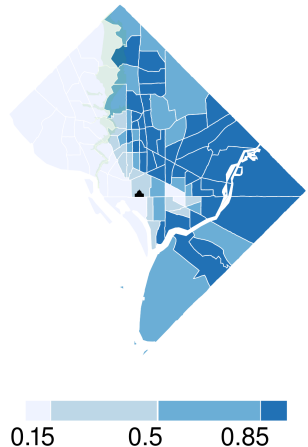
Profound Changes: Share African American by Neighborhood

1930
1940
1950
1960
1970
1980
1990
2000
2010



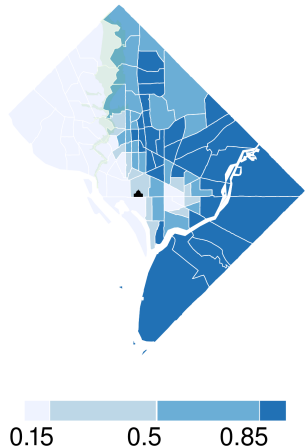
Profound Changes: Share African American by Neighborhood

1930
1940
1950
1960
1970
1980
1990
2000
2010



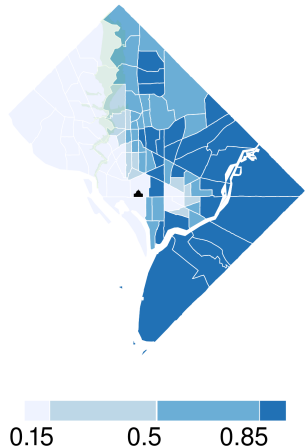
Profound Changes: Share African American by Neighborhood

1930
1940
1950
1960
1970
1980
1990
2000
2010



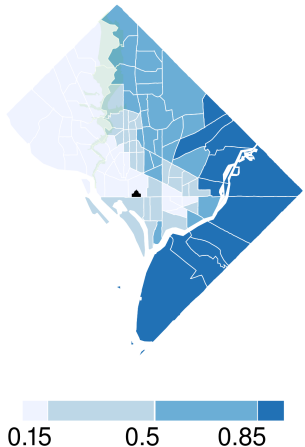
Profound Changes: Share African American by Neighborhood

1930
1940
1950
1960
1970
1980
1990
2000
2010



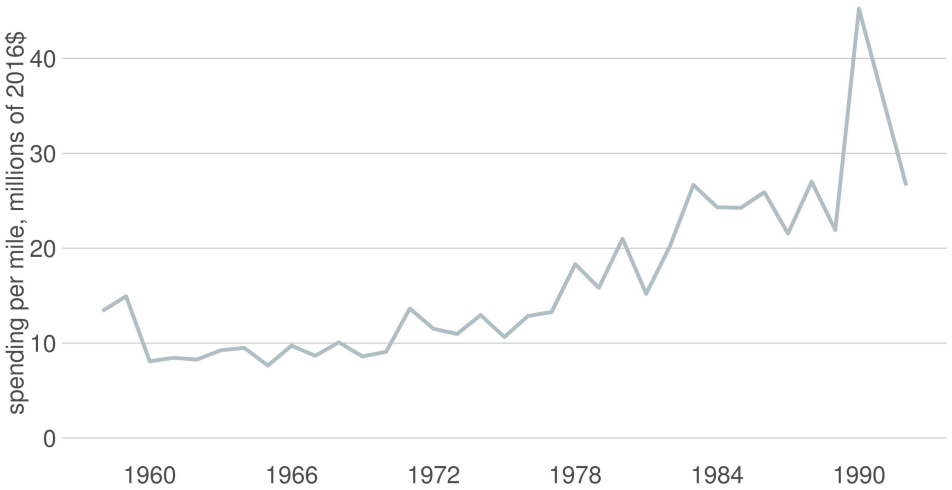
Profound Changes: Share African American by Neighborhood

1930
1940
1950
1960
1970
1980
1990
2000
2010

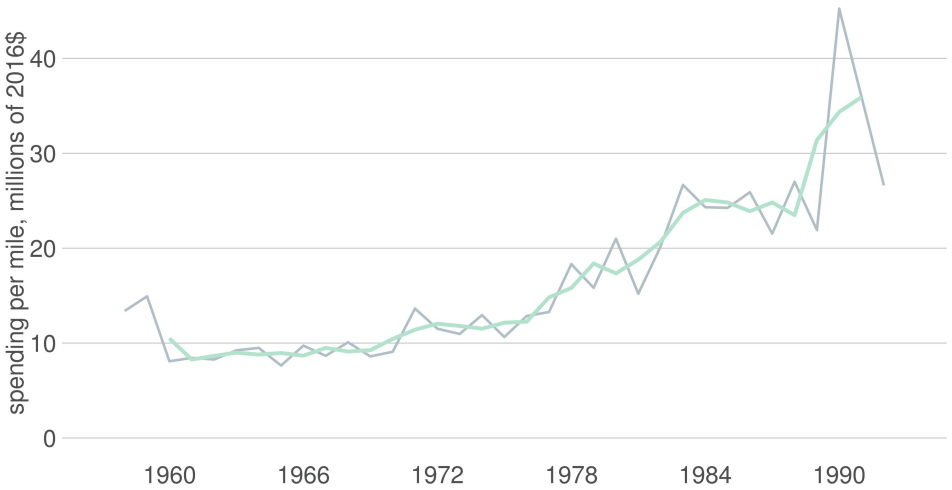


From a project about whether and why infrastructure costs are increasing

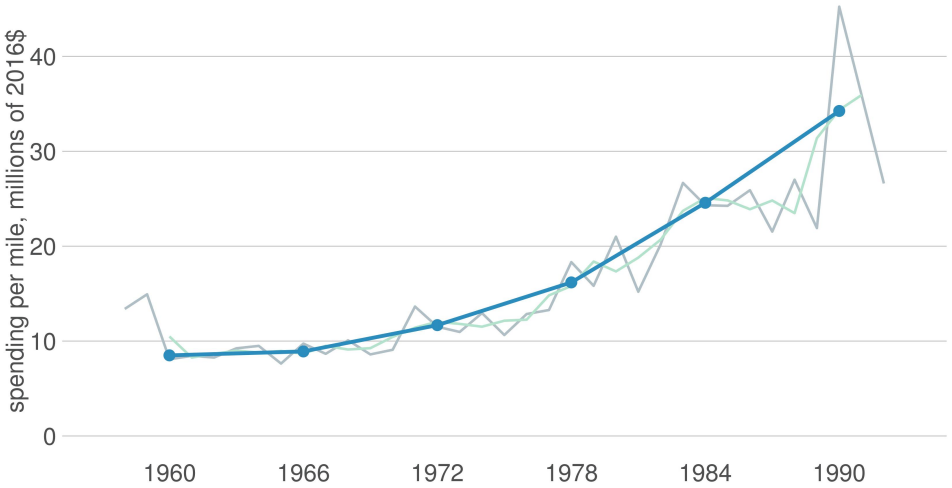
Spending Per Mile has Tripled Since 1960s



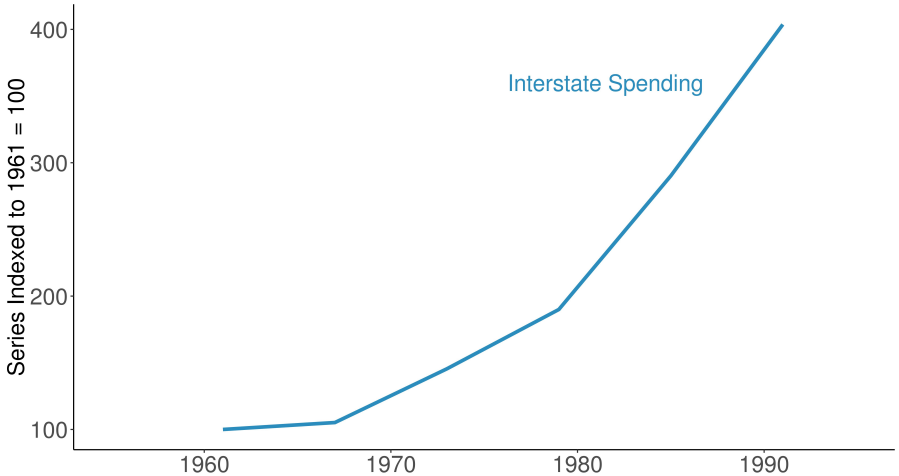
Spending Per Mile has Tripled Since 1960s



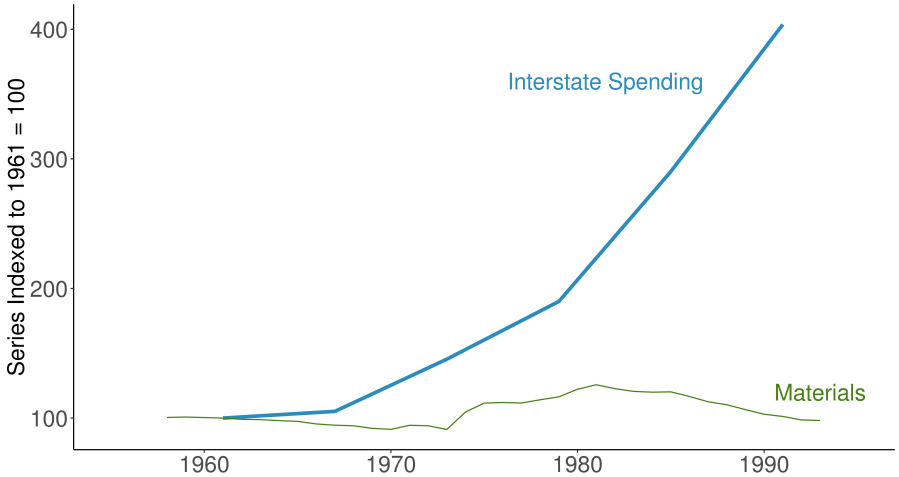
Spending Per Mile has Tripled Since 1960s



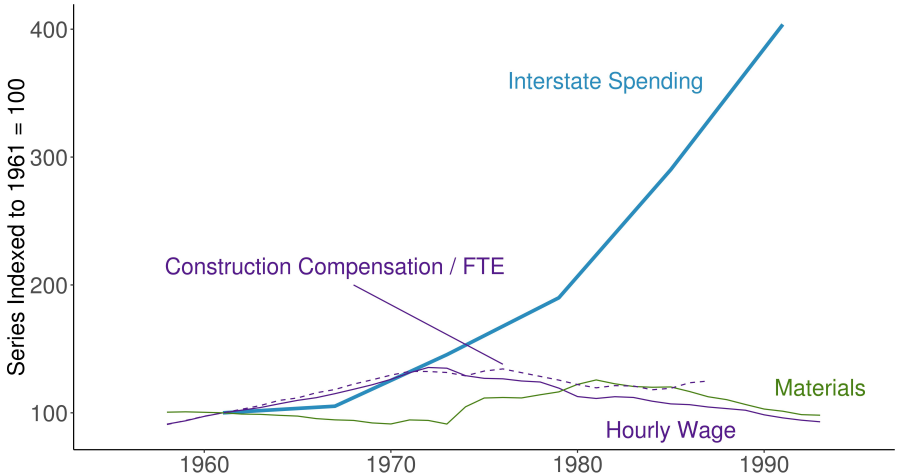
Interstate Spending Per Mile, Indexed to 100 in 1961



Materials Prices are Roughly Flat Over the Period

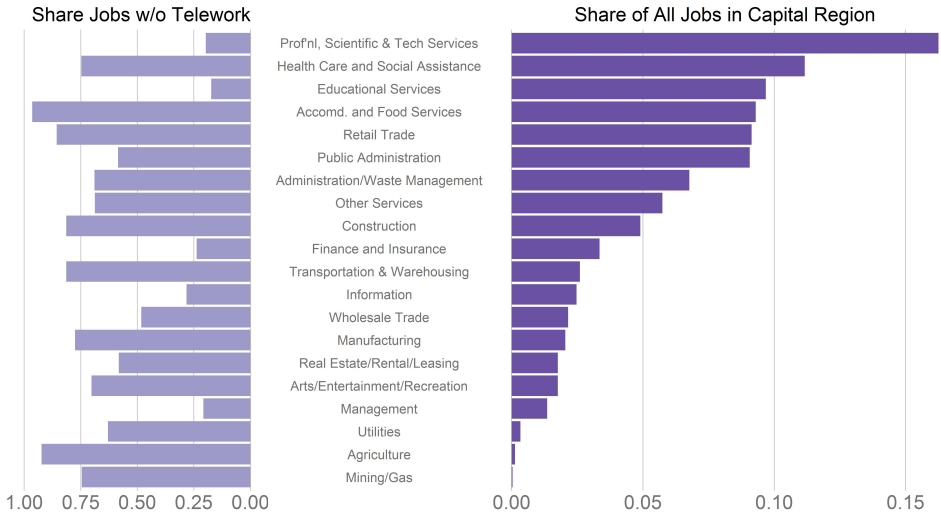


Wages Are Flat, Too → Input Prices Cannot Explain Increase

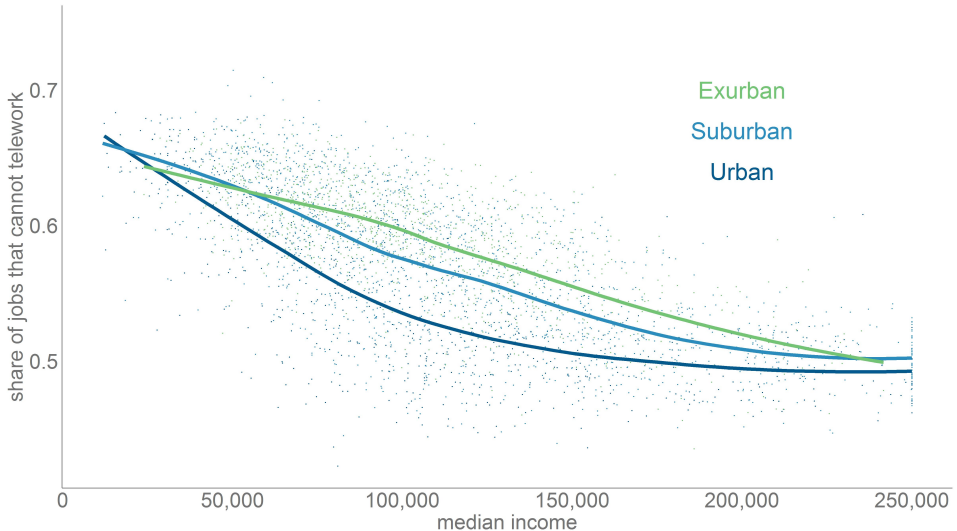


From a project about the likely impacts of Covid on the DC region

Capitol Region Strong in Work-from-home Sectors



Wealthier People More Likely to Be Able to Telework



Tufte

Tufte

1. Why Tufte?
2. Beginnings of graphics
3. Why visualizations help
4. Tufte's four types of graphs, with examples
5. Tufte's problems with graphics
6. Rules of graphic integrity

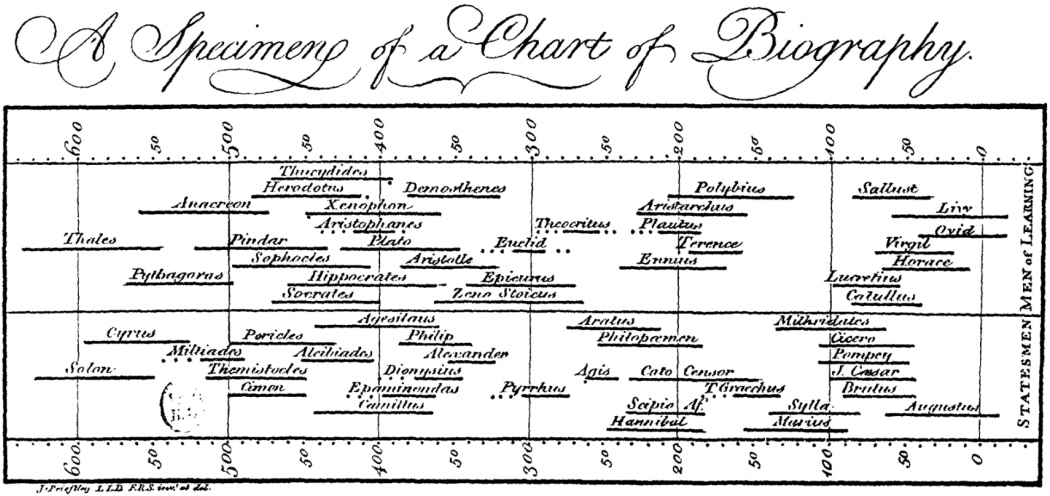
Edward Tufte

- A quantitative political scientist
- Writing in the mid-1970s
- Became interested in visualization by working with pioneering statistician John Tukey
- Remember that this is the pre-Excel era, in which data graphics are difficult to make

Why Do We Read This?

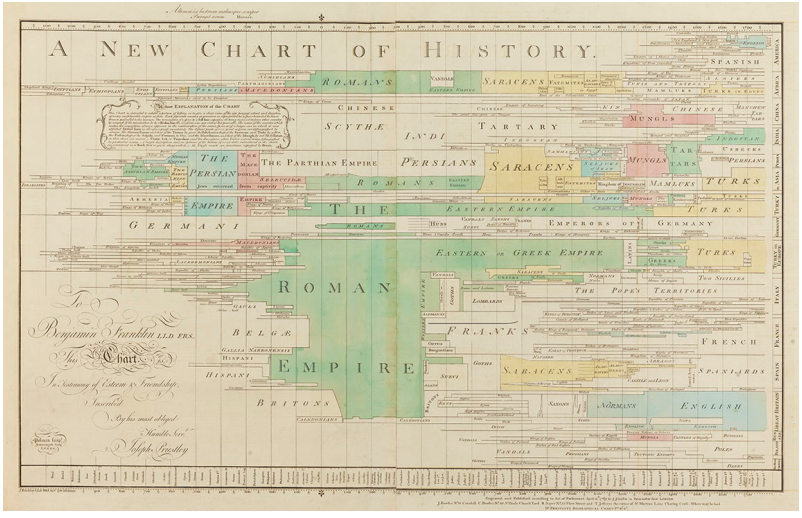
- Among the first to take the field as a whole seriously
- Greatest popularizer of a now-accepted set of conventions
- Highlights that visualizations only began
  - 1765 with Joseph Priestley
  - 1786 with William Playfair

Priestely's Sensation



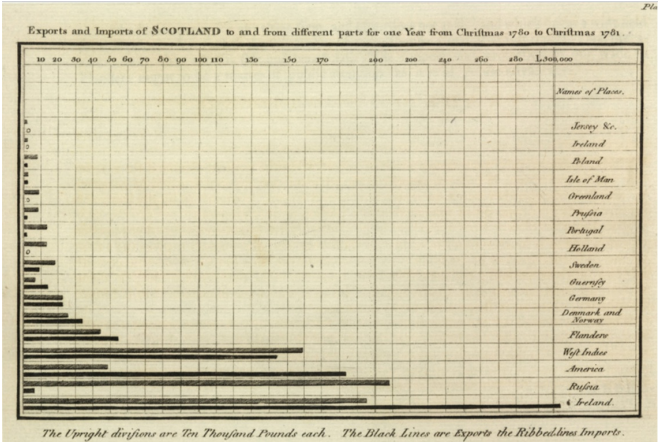
Joseph Priestley (1733-1804), chart from 1765. [Public domain via Wikipedia]. See discussion here.

And a Full Scale Version of the Same



Joseph Priestley (1733-1804), chart from 1769, Library Company of Philadelphia. Source here.

The World's First Bar Chart



William Playfair (1759-1823), 1786. [Public domain via Wikipedia]

An Argument for Better Visualization

Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

All series have the same

- mean of X
- variance of X
- mean of Y
- variance of Y
- $\text{corr}(X, Y)$
- $\hat{\beta}$
- R^2

An Argument for Better Visualization

Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

All series have the same

- mean of X
- variance of X
- mean of Y
- variance of Y
- $\text{corr}(X, Y)$
- $\hat{\beta}$
- R^2

Which one is a vertical line?

An Argument for Better Visualization

Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

All series have the same

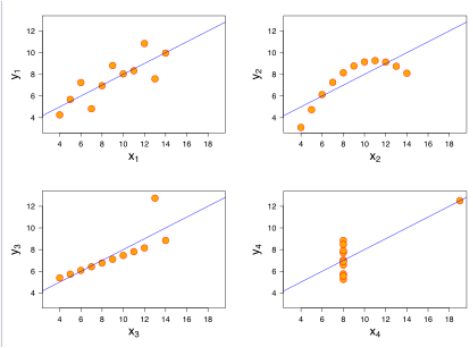
- mean of X
- variance of X
- mean of Y
- variance of Y
- corr(X, Y)
- $\hat{\beta}$
- R^2

Which one is a vertical line? Which one is an upside-down U?

Thanks to Wikipedia for quartet table.

An Argument for Better Visualization

Because good visualizations tell the most compelling story



Thanks for Wikipedia for figure.

Tufte's Types of Graphs

1. Data maps
2. Time series
3. Space-time narrative designs
4. Relational graphs – the holy grail

Data Maps

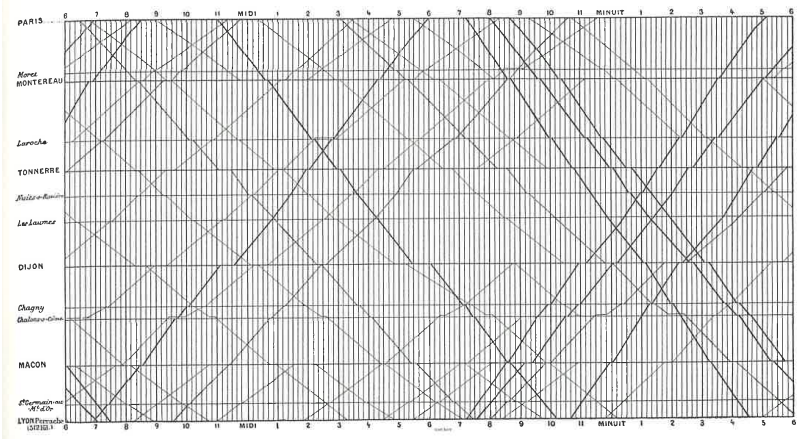
- Describe the location of numbers
- This can be revealing or obfuscating
- We will make these in this class
- A product of the mid-1800s

John Snow on the Location of Cholera in London, c. 1850

Time Series

- Time on the horizontal axis
- Something else on the vertical axis
- One of the first types of data graphics

Train, Paris to Lyon

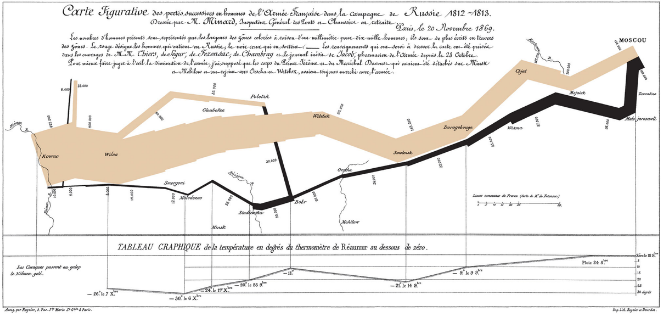


See Tufte for citation.

Space-Time Narrative Designs

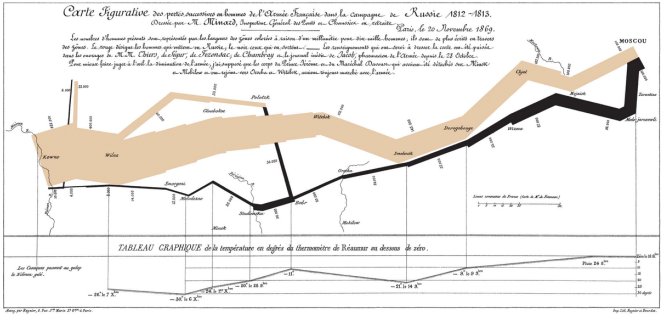
- Move over space and time at the same time
- A time series plus

Space-Time Narrative Example



Which dimensions?

Space-Time Narrative Example



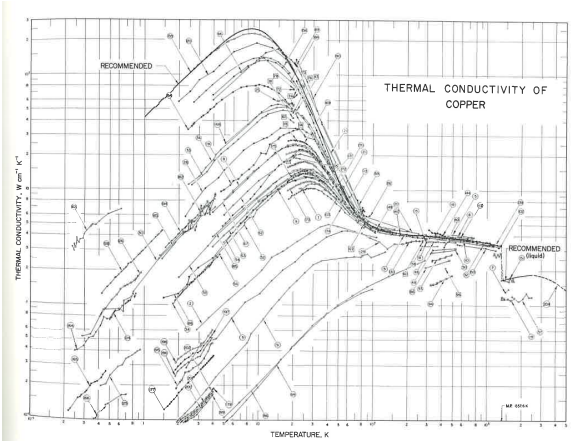
Which dimensions?

1. army size
2. army location, N/S
3. army location, E/W
4. direction of movement
5. temperature
6. by date

Relational Graphics

- One variable on the vertical, another on the horizontal
- A conceptual advance in graphics
- A more sophisticated way of thinking

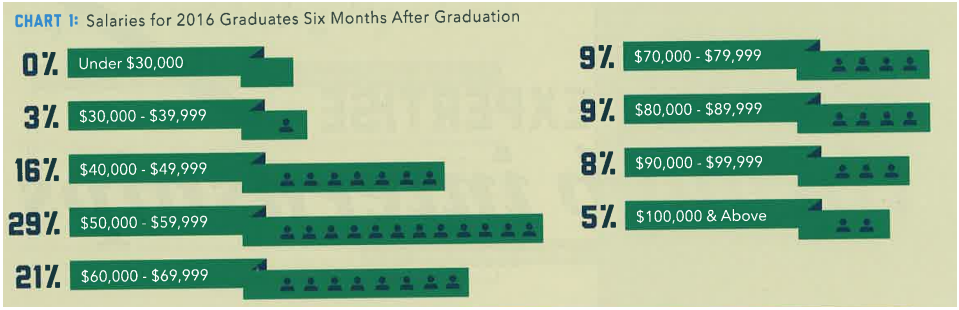
Relational Graphics Example



Tufte's Main Causes of Distortion in Graphics

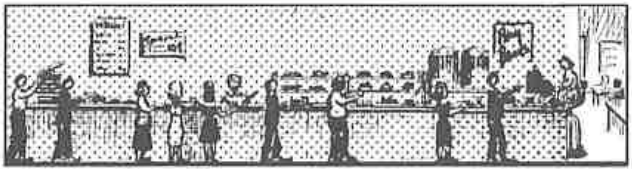
1. Data are bad
   - should be per capita and are not
   - data are not consistent over time
   - don't adjust for inflation
2. Graphics are rotten
   - size doesn't match the numbers
   - colors and styles are misleading
   - graphic fails to highlight key point
3. Graphics are irrelevant
   - too much extraneous stuff

Ex. of 2: Size and Number Don't Match



Ex. of 3: Graphics are Irrelevant

The Company Cafeteria was used by 9 Out of 10 Employees during the Fiscal Year 1949



0 100%

Source: COMPANY REPORTS

Tufte's Six Rules of Graphic Integrity, 1 to 3 of 6

1. The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented.
2. Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data.
3. Show data variation, not design variation.

Tufte's Six Rules of Graphic Integrity, 4 to 6

4. In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units.
5. The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.
6. Graphics must not quote data out of context.

Getting Started with R

What is R?

- A programming language
- Developed by statisticians from New Zealand
- Open source, and therefore free
- Based on "S," developed by Bell Labs

Strengths of R

- Free
- Open-source, so packages by all kinds of users are available
- There are frequently many ways to do the same task
- Very good graphics
- Very flexible
- Can have many datasets in memory at once
- Can analyze large datasets
- Can do maps **and** spatial analysis
- Big user community and lots of online help

Weaknesses of R

- Not always enterprise-ready: packages break and there is no central help
- There are frequently many ways to do the same task
- Syntax can be challenging
- Syntax is inconsistent across packages

Today's Goals

- Digest info and ask questions to me about R
- When you finish today's tutorial, you will be able to
  - run a R script
  - create a R dataframe
  - do basic operations with a R dataframe
- Please work together! Now and later
- And turn in **your own work in your own words**

Today's Goals

- Digest info and ask questions to me about R
- When you finish today's tutorial, you will be able to
  - run a R script
  - create a R dataframe
  - do basic operations with a R dataframe
- Please work together! Now and later
- And turn in **your own work in your own words**
- Feel free to lean on ChatGPT and the like – let me know how it goes
- Cite when you use AI, making clear which are not your own words

R Tools

Rest of Today's Class

- Link to today's tutorial from lectures page
- Make sure you can do the `hello_world.R` program before leaving
- You'll continue work at home on your own and turn in a problem set next lecture

Next Lecture: Lecture 2

- Turn in work for tutorial 1
- Read Few Chapters 3 and 5
- Look at "Graph Choice Chart"