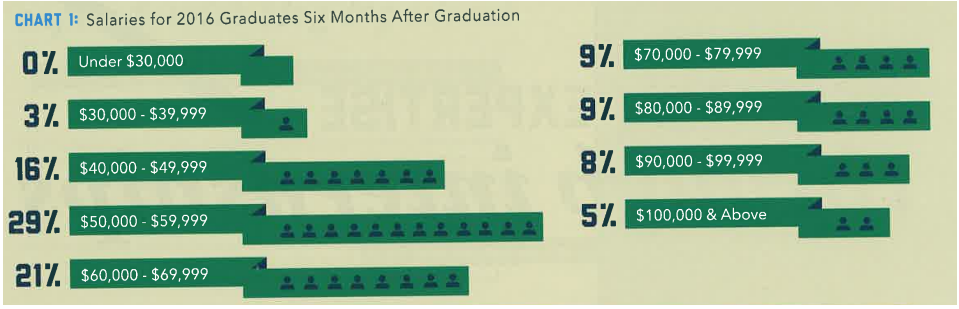
Lecture 1:
Welcome to Data Visualization Using R

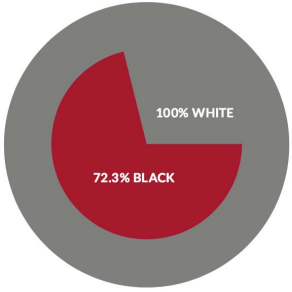
January 13, 2020

Take This Class So You Won't Make This Graphic



From Trachtenberg's 2018 magazine.

Or This One

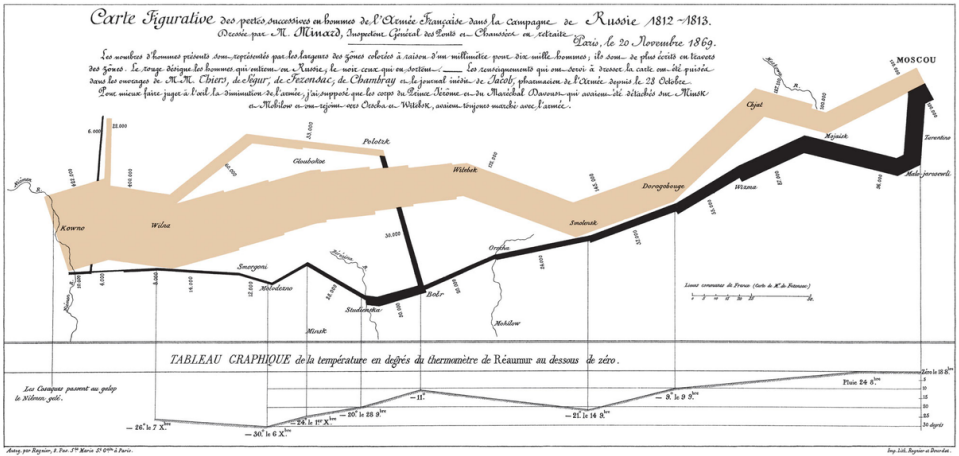


EQUALITY INDEX OF BLACK AMERICA, 2016-2017

	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

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

Overview

Course Administration

Some R Examples

Tufte, Grandfather of Visualization

Getting Started with R

R Programming

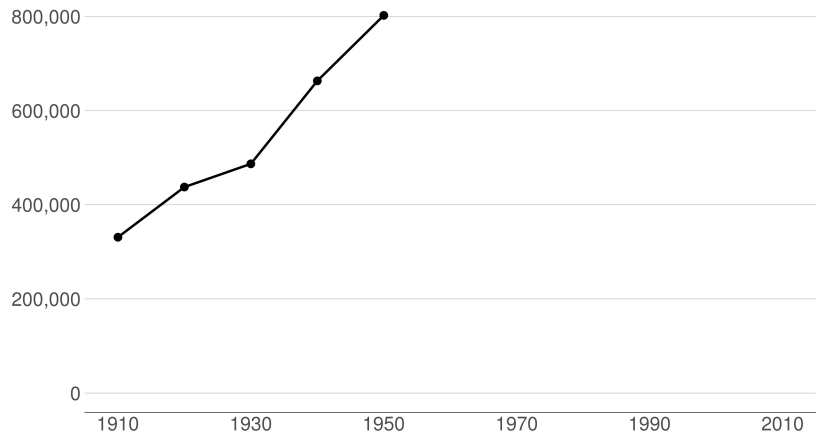
Course Administration

1. Syllabus
   - Policy brief handout
   - Fully composed chart handout
   - Good/bad/ugly assignments handout
2. Bring a name tent to class
3. Questions/issues with readings?
4. Make sure you're signed up for Piazza
5. Introductions
   - name and degree
   - why this course?
   - what you do now
   - what you'd like to do when you're done

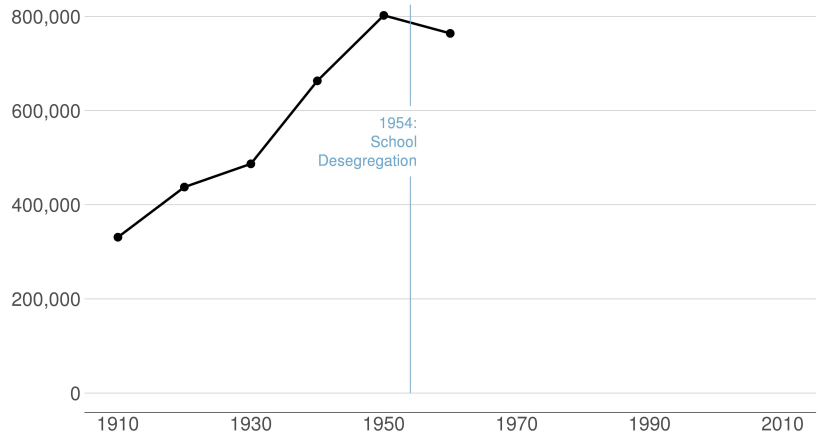
R Examples

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

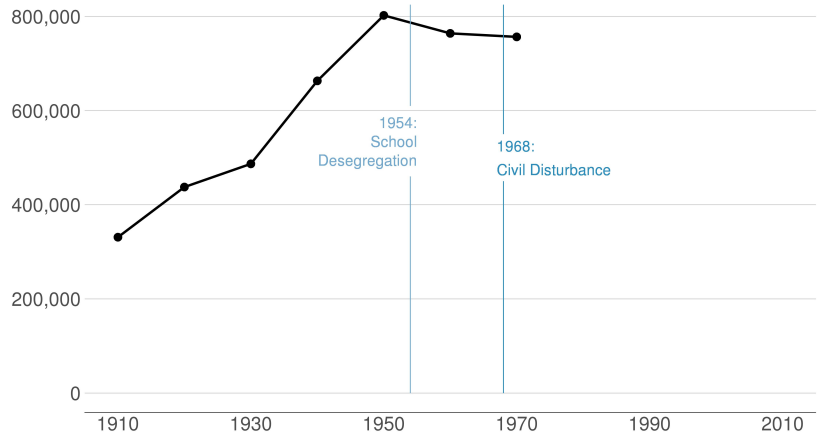
DC Gains Population Through 1950



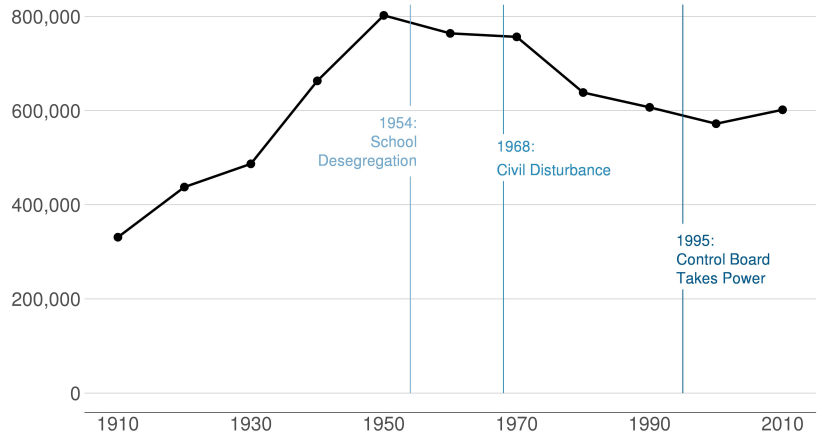
Population Loses Start with Desegregation



Continue After Civil Disturbance

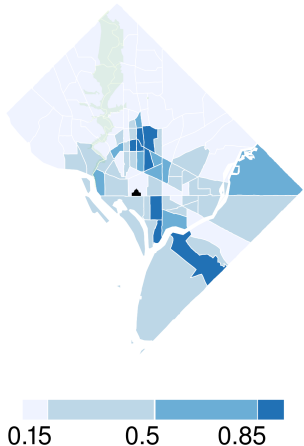


Population Turns Up After 2000



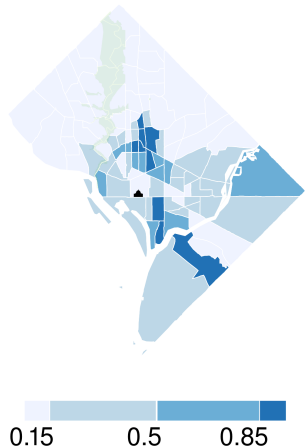
Profound Changes: Share African American by Neighborhood

1930
1940
1950
1960
1970
1980
1990
2000
2010



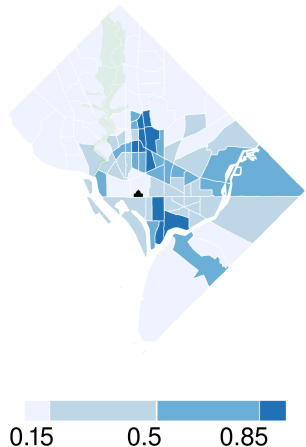
Profound Changes: Share African American by Neighborhood

1930
1940
1950
1960
1970
1980
1990
2000
2010



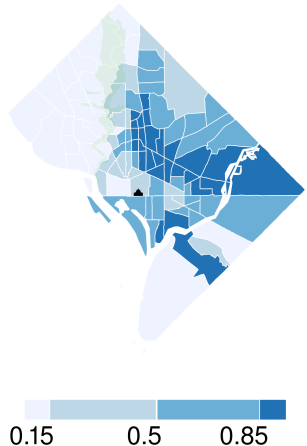
Profound Changes: Share African American by Neighborhood

1930
1940
1950
1960
1970
1980
1990
2000
2010



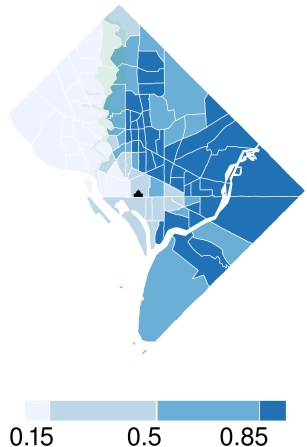
Profound Changes: Share African American by Neighborhood

1930
1940
1950
1960
1970
1980
1990
2000
2010



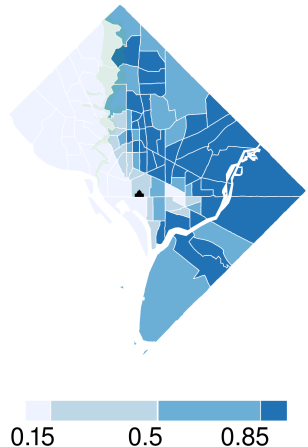
Profound Changes: Share African American by Neighborhood

1930
1940
1950
1960
1970
1980
1990
2000
2010



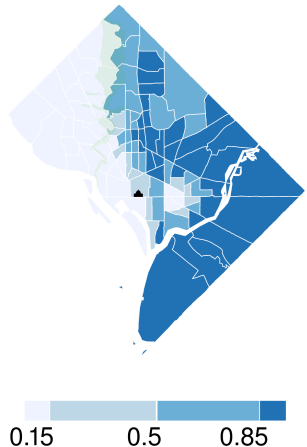
Profound Changes: Share African American by Neighborhood

1930
1940
1950
1960
1970
1980
1990
2000
2010



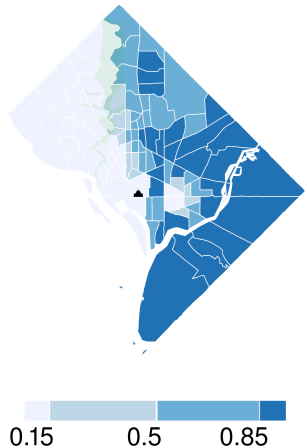
Profound Changes: Share African American by Neighborhood

1930
1940
1950
1960
1970
1980
1990
2000
2010



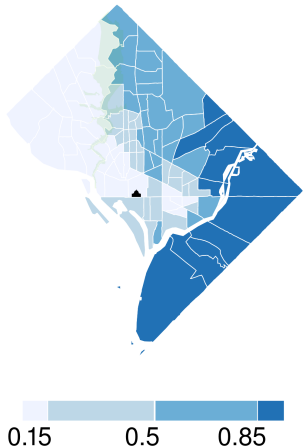
Profound Changes: Share African American by Neighborhood

1930
1940
1950
1960
1970
1980
1990
2000
2010

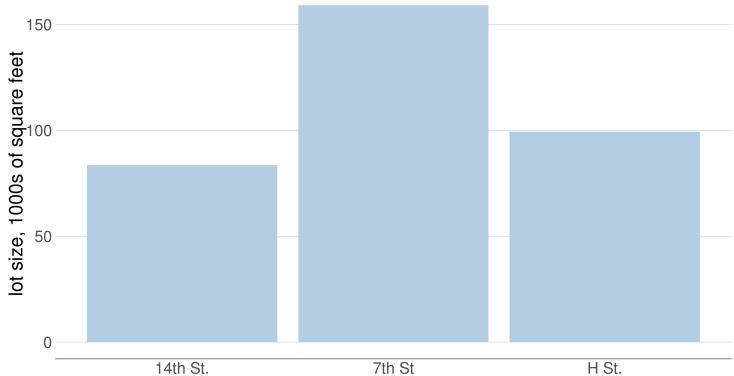


Profound Changes: Share African American by Neighborhood

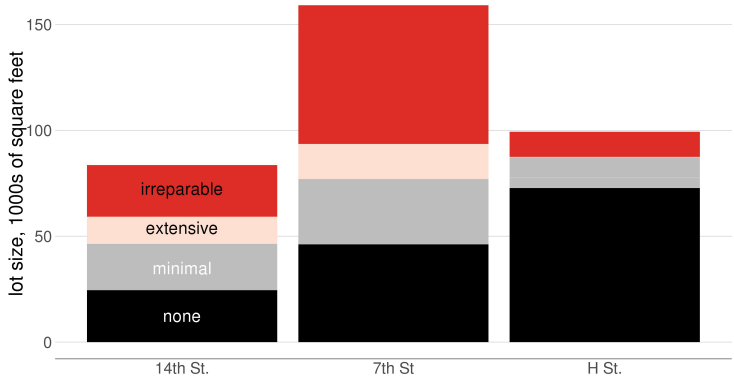
1930
1940
1950
1960
1970
1980
1990
2000
2010



By Square Footage, 7th Street is Most Impacted

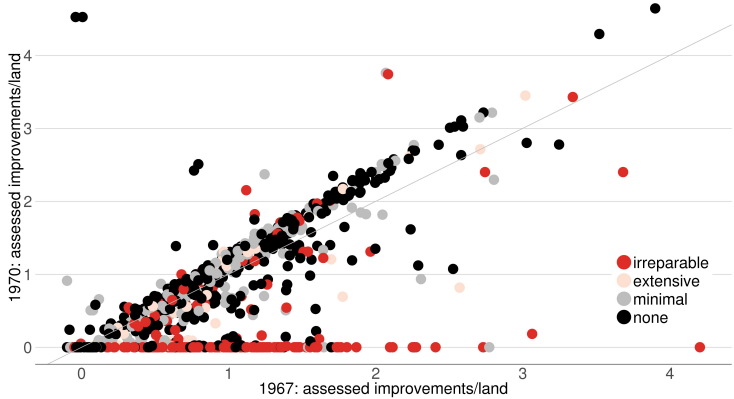


Roughly Half of Square Footage Damaged to Some Degree

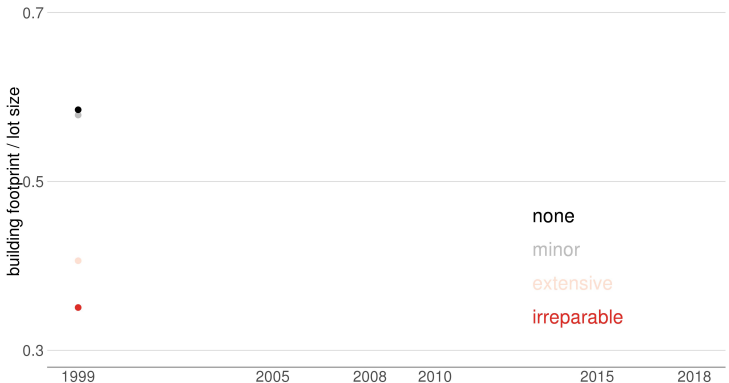


Assessed Value of Most Improvements Drops, 1967 to 1970

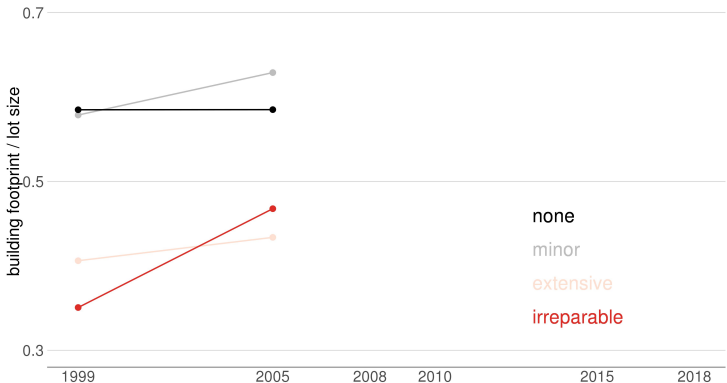
Damaged Properties Lose Improvements, A Few Rebuild



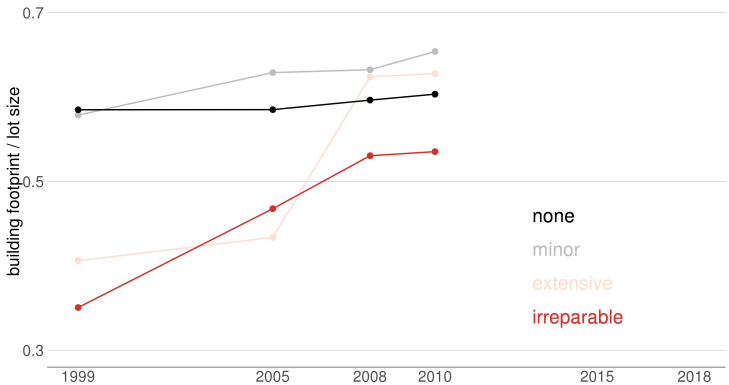
1999: Damaged Properties Have Smaller Structures



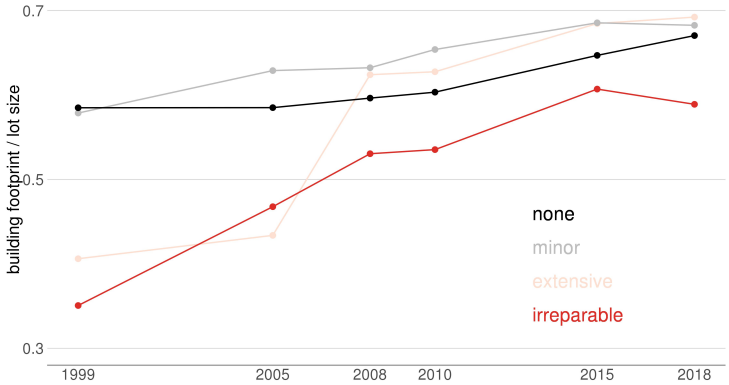
2005: Damaged Properties Show Some Catch-up



2010: Damaged Properties Approaching Undamaged Ones

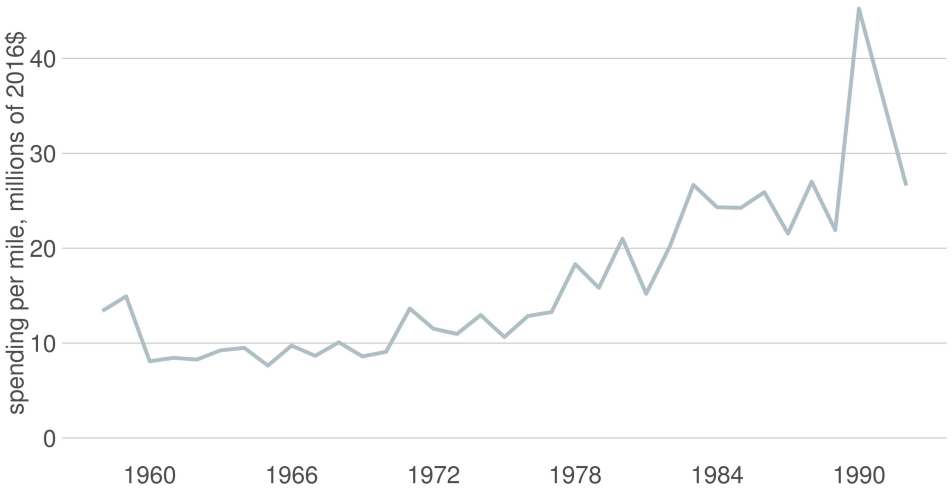


2018: Near Convergence of Damaged Properties

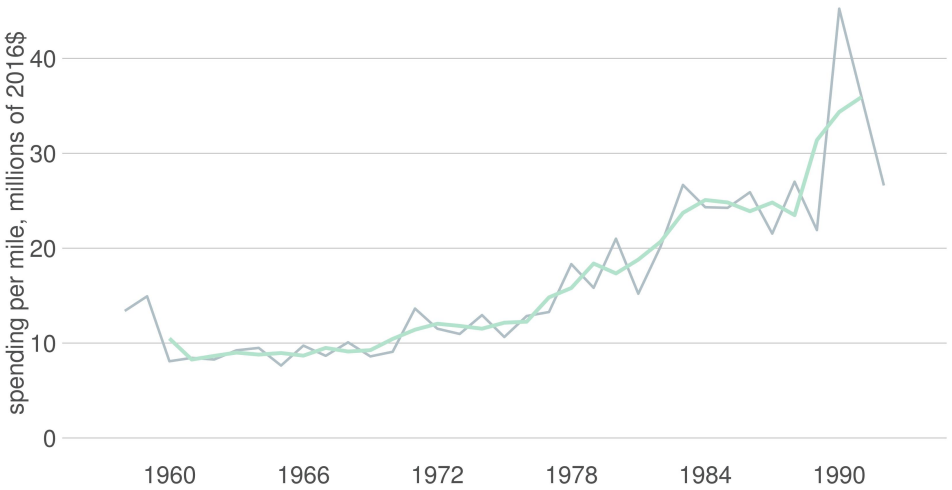


From a project about whether and why infrastructure costs are increasing

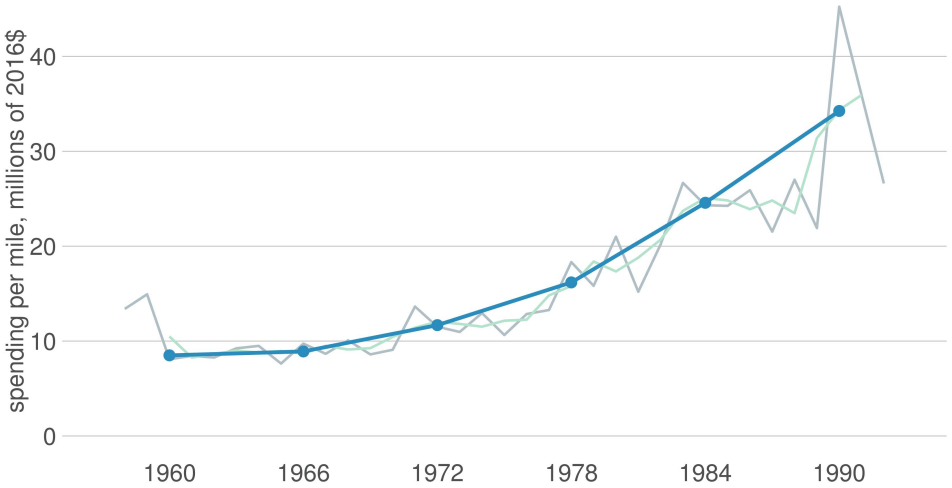
Spending Per Mile has Tripled Since 1960s



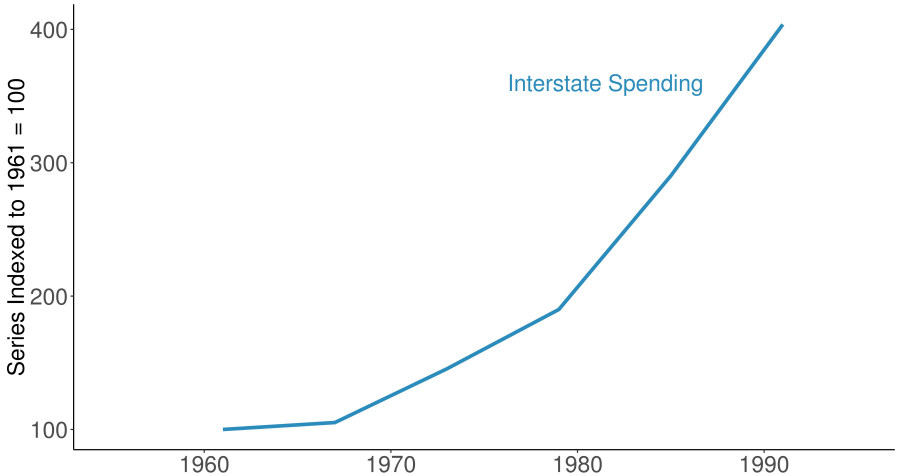
Spending Per Mile has Tripled Since 1960s



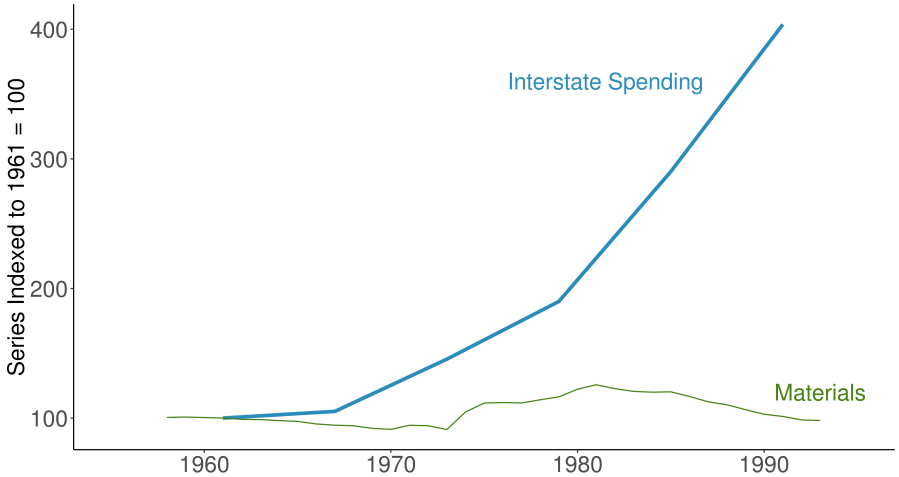
Spending Per Mile has Tripled Since 1960s



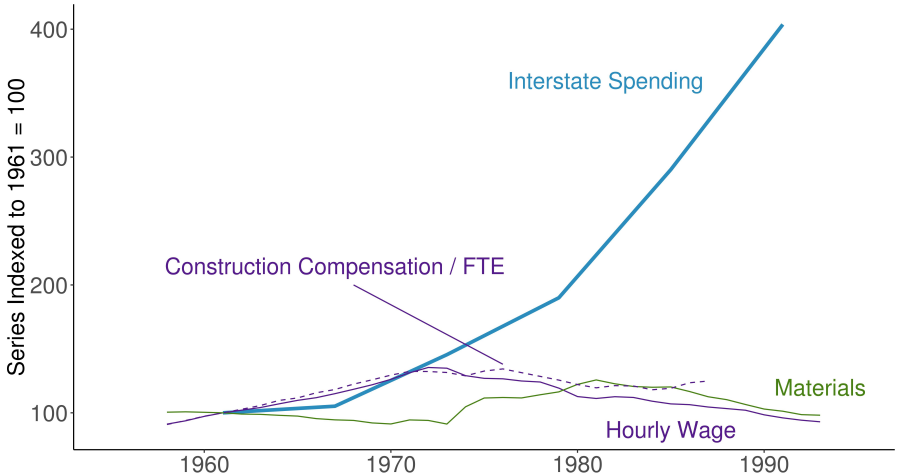
Interstate Spending Per Mile, Indexed to 100 in 1961



Materials Prices are Roughly Flat Over the Period



Wages Are Flat, Too → Input Prices Cannot Explain Increase



Tufte

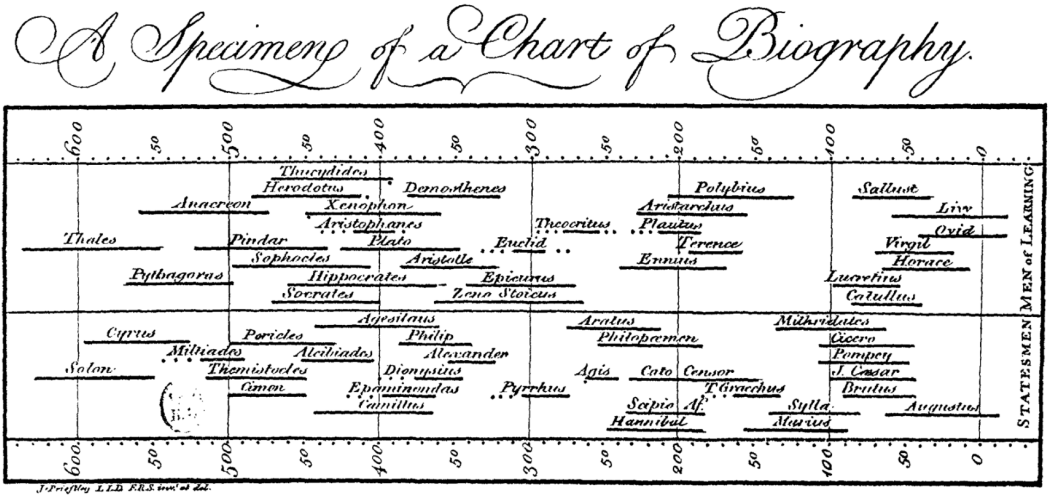
Edward Tufte

- 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?

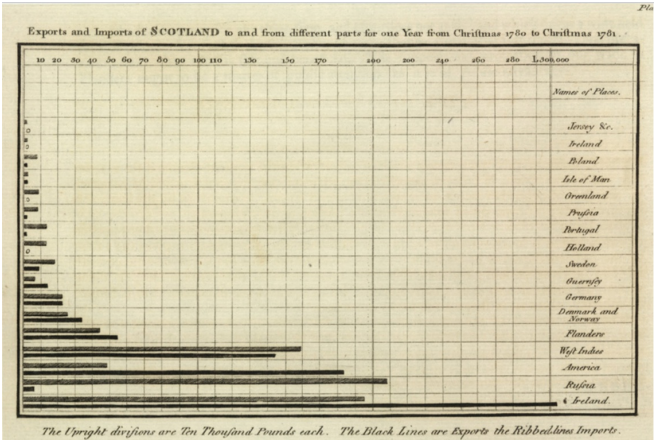
- Among the first to take the field as a whole seriously
- Greatest popularizer of a now-accepted set of conventions
- Highlights that visualizations only began
  - 1765 with Joseph Priestley
  - 1786 with William Playfair

Priestly's Sensation



Joseph Priestley (1733-1804) [Public domain via Wikipedia]

The World's First Bar Chart



William Playfair [Public domain via Wikipedia]

An Argument for Better Visualization

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

An Argument for Better Visualization

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?

An Argument for Better Visualization

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

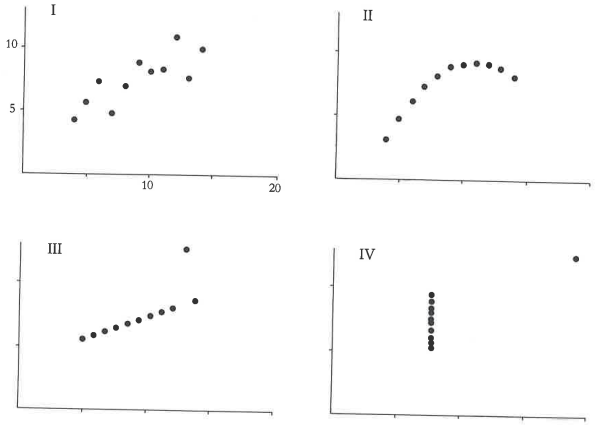
All series have the same

- 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?

An Argument for Better Visualization

Because good visualizations tell the most compelling story



Tufte's Types of Graphs

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

Data Maps

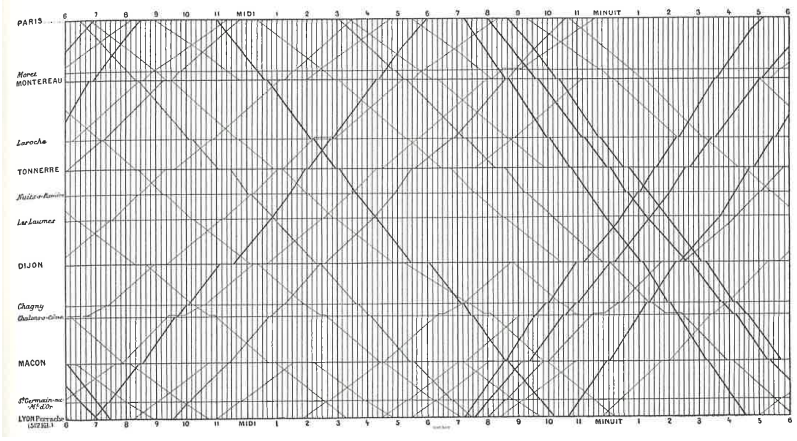
- Describe the location of numbers
- This can be revealing or obfuscating
- We will make these in this class
- A product of the mid-1800s

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

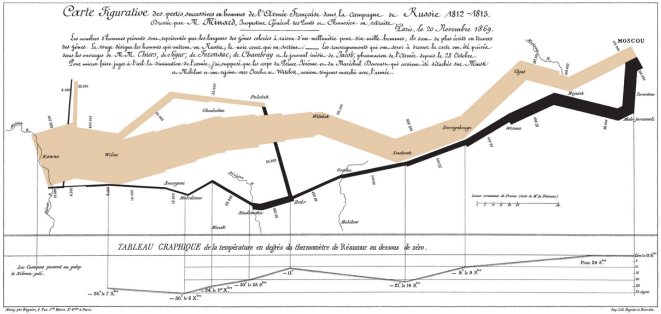


See Tufte for citation.

Space-Time Narrative Designs

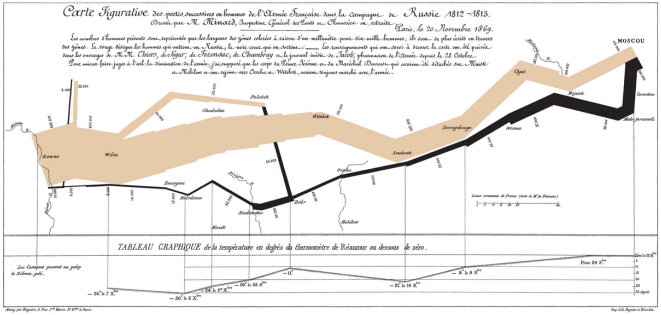
- Move over space and time at the same time
- A time series plus

Space-Time Narrative Example



Which dimensions?

Space-Time Narrative Example



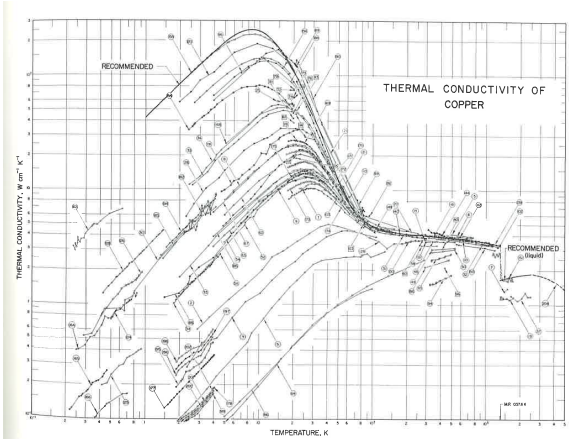
Which dimensions?

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

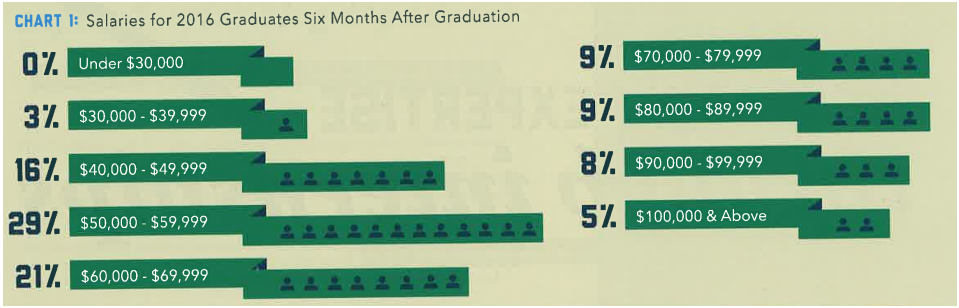
Relational Graphics Example



Tufte's Main Causes of Distortion in Graphics

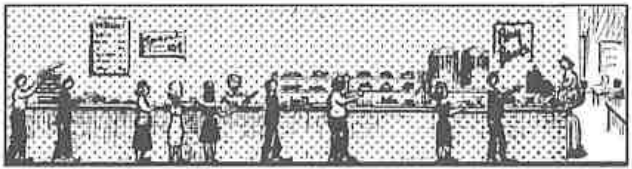
1. Data are bad
   - should be per capita and are not
   - data are not consistent over time
   - don't adjust for inflation
2. Graphics are rotten
   - size doesn't match the numbers
   - colors and styles are misleading
   - graphic fails to highlight key point
3. Graphics are irrelevant
   - too much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

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

Today's Goals

- When you leave today, you will be able to
  - run a R script
  - create a R dataframe
  - do basic operations with a R dataframe
- Download the R tutorial for this class now.
- You'll continue work at home on your own and turn in a problem set next lecture

R Tools

Today

A. Hello World

B. A R dataframe

C. Packages

D. Subsetting

E. Functions

F. Summarizing

A. Hello World

- the very first computer program prints “Hello World”
- so we start with this

```
print("hello world!")
```

A. Hello World

- the very first computer program prints “Hello World”
- so we start with this

```
print("hello world!")
```

```
## [1] "hello world!"
```

A. Hello World v.2

- make an object that holds the value “hello world”
- print that object

```
mr.object <- "hello world"
mr.object
```

```
## [1] "hello world"
```

B. A R dataframe

- a dataframe is the basic building block of data analysis in R
- R has other types of data structures, but this will the be most useful to you
- dataframe consists of columns
- each column can be
  - numeric: 1,2,3.556,-2.6
  - or
  - character: “hello”,“dogs are good”,""
- refer to rows and columns

Sample dataframe

```
new.dataframe <-
  data.frame(class = c(1,2,3),
             subject = c("basics","merging","graphs"),
             students = c(19,19,18))
new.dataframe
```

```
##   class subject students
## 1     1  basics       19
## 2     2 merging       19
## 3     3  graphs       18
```

Referring to parts of the dataframe

```
new.dataframe[ROWS,COLUMNS]
```

Referring to parts of the dataframe

```
new.dataframe[ROWS,COLUMNS]
```

Just one column, all rows

```
new.dataframe[,c("students")]
```

```
## [1] 19 19 18
```

Referring to parts of the dataframe

```
new.dataframe[ROWS,COLUMNS]
```

Just one column, all rows

```
new.dataframe[,c("students")]
```

```
## [1] 19 19 18
```

Just two rows, all columns

```
new.dataframe[1:2,]
```

```
##   class subject students
## 1     1  basics       19
## 2     2 merging       19
```

Refer to just one column with dollar sign

- you can also refer to one specific variable as

```
new.dataframe$students
```

C. Packages

- there is “Base R,” which is a set of basic commands
- and user-written packages that add functionality
- some packages are maintained by teams, frequently updated, and do many things
- some are one-function add-ins
- most famous are those by Hadley Wickham
- today we’ll use his “dplyr” packcage

Installing packages

- install a package once

```
install.packages("dplyr", dependencies = TRUE)
```

Installing packages

- install a package once

```
install.packages("dplyr", dependencies = TRUE)
```

- call a package at the beginning of any program in which you'd like to use the package

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

D. Subsetting

- sometimes you want to work with something smaller than the whole dataframe
- create a new dataframe that has only part of the big one
- here we keep just students 1 and 2

```
df.smaller <- new.dataframe[1:2,]
df.smaller
```

```
##   class subject students
## 1     1  basics       19
## 2     2 merging       19
```

Subset by attributes

- take only classes with 19 students

```
df19 <-
  new.dataframe[which(new.dataframe$students == 19),]
df19
```

```
##   class subject students
## 1     1  basics       19
## 2     2 merging       19
```

E. Functions

- R has 1000s of functions
- functions take data and do something to it
- general format is

```
new.output <- function(inputs)
```

where `inputs` can be a dataframe or something else

The Mean Function

- suppose we want to know the average number of students
- use the mean function

```
mean(x = new.dataframe$students)
```

```
## [1] 18.66667
```

The Mean Function

- suppose we want to know the average number of students
- use the mean function

```
mean(x = new.dataframe$students)
```

```
## [1] 18.66667
```

or

```
new.mean <- mean(x = new.dataframe$students)
new.mean
```

```
## [1] 18.66667
```

F. Summarizing

- frequently, you'd like to know something at a level of aggrgation not in your dataset
- in our case, maybe average attendance
- make a new dataframe with this information
- use `dplyr` library

Making a new dataset that is a function of the old one

```
av.attendance <-
  summarize(.data = new.dataframe,
            av.at = mean(students, na.rm = TRUE))
av.attendance
```

```
##       av.at
## 1 18.66667
```

- more complicated example in tutorial

Next Lecture

- Turn in PS 1, which is at the end of the tutorial
- Read Few Chapters 3 and 5
- Look at "Graph Choice Chart"