Lecture 3:
Bar Graphs

February 1, 2021

Overview

Course Administration

Good, Bad and Ugly

General Graph Design, Few Ch. 9

What is a Bar Chart?

ggplot and Histograms

Course Administration

1. Our TA is hired! Kim Wilson is here

Course Administration

1. Our TA is hired! Kim Wilson is here
2. Reading quiz

Course Administration

1. Our TA is hired! Kim Wilson is here
2. Reading quiz
3. Hopefully your policy brief proposal is in – I will try to return by next week
4. Make sure you're checking Piazza
5. Anonymous google sheet with grades up in a few weeks
6. List of R commands by lecture: see Resources tab
7. Consider installing RMarkdown for problem set submission

Feedback from Tutorial 1

- Need one **pdf** that contains
  - Easy-to-read questions with answers – not raw output
  - R code, tutorial and problem set
  - R output – copied from console is fine
- Q4: "find state-level (or the part of the state we observe) average population over the entire period"
  - this means find state-level population: add up to the state level (just for DC-area jurisdictions)
  - then average over all years

Next Week's Good Bad and Ugly

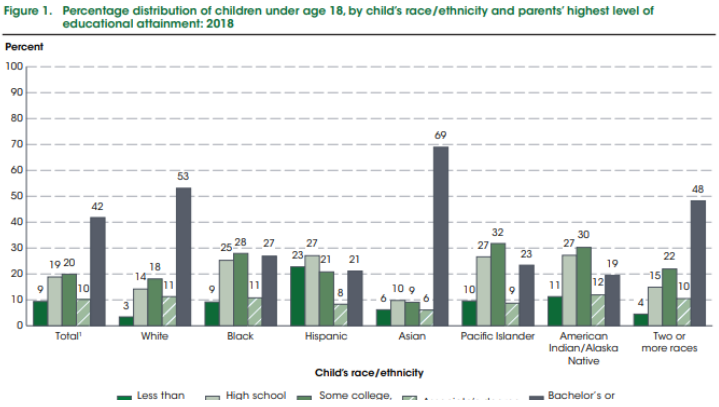
Post by Wednesday noon. You link on google sheet. Earlier is ok.

Finder	Commenter
Trinady M.	Sarah H.
Eleanor T.	Preston B.

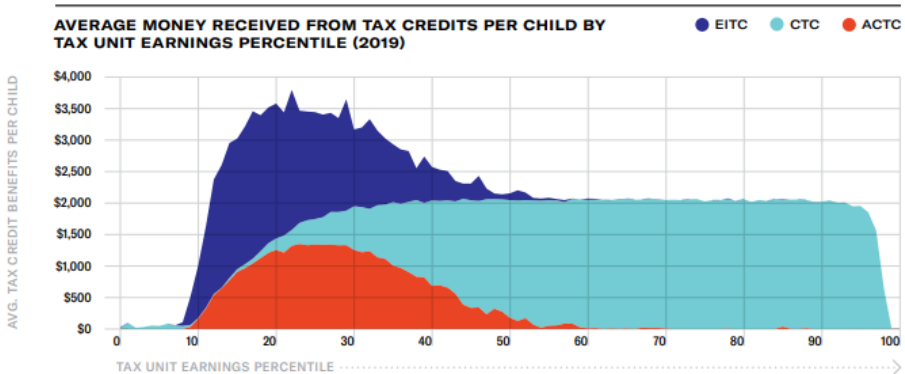
This Week's Good Bad and Ugly

Gabriel and Anthony

Gabriel Has the Floor



Anthony Has the Floor



Online Portion of Lecture 3

To in-person lecture

Few: Three Primary Ways to Convey Information

1. Form
2. Color
3. Spatial Position

Leading to Two Key Rules for All Graphics

1. Maintain visual correspondence to quantity
2. Avoid 3D

GRAPH CHOICE CHART

Does your question ask you...

about the **variability** of a group of data points? (i.e. the range of the data, the shape of the distribution, or what the center of the data is)

1. Do all high tides rise to the same height?
2. How variable are wind speeds in Denmark?
3. What is the range and distribution of incomes in Sudan?

→ Frequency Plot → MAKE EITHER → Histogram, Box Plots, Dot Plot

to **compare two or more groups** to decide if the groups are the same or different?

Do you want to compare the **variability of all data points** in each group to decide if any difference between the groups is meaningful?

1. Which of the two solar cars consistently goes the farthest?
2. Is there a meaningful difference in the heights of fertilized and unfertilized bean plants?

→ FOR EACH GROUP MAKE A → Frequency Plot → MAKE EITHER → Box Plots, Dot Plot

Are you comparing **single numbers** that summarize a group? (such as mean, median, or total...)

1. Was the total snowfall greater this winter than last winter?
2. Do cats and dogs have the same average body temperature?
3. How do the median incomes for the US and India compare?

→ Bar Graph

if **two numeric factors are correlated**?

1. Is the temperature inside the house correlated with the temperature outside?
2. How did electricity used by the kitchen circuit fluctuate during the past week?

Does it ask about how something changes through **linear TIME**?

N
1. Is the fuel efficiency of a car related to its weight?
2. Are smoking rates correlated with median income?
3. Given a fixed volume, how are temperature and pressure related?

→ Scatter Plot

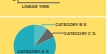
Y
1. Is sea level rising?
2. How did my weight change over the last 3 months?

→ Line Graph

how a **total** is **proportioned** into sub-groups? (Or what proportion a sub-group is of a total?)

1. What were Brazil's most significant exports in 2015?
2. What proportion of global electricity production comes from wind?
3. How do Parisians typically commute to work?

→ MAKE EITHER → Pie Chart, Stacked Bar Chart



V0.1 updated 3.29.16

Bar Charts

Bars Outline

Big idea: relative size

Bars Outline

Big idea: relative size

- What do bars do?
- Few, Ch. 10, bits of bars
- Lollipops, esp from WSJ
- Giant numbers from WSJ

What Does a Bar Chart Do?

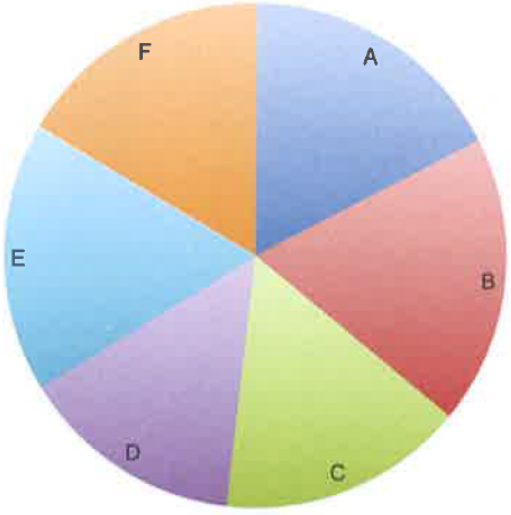
Bars compare quantities across categories

What Does a Bar Chart Do?

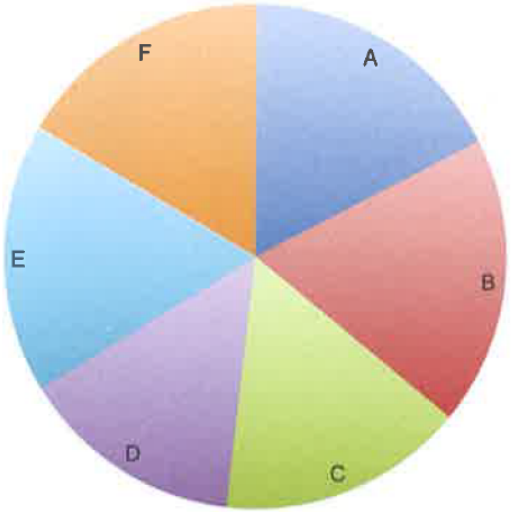
Bars compare quantities across categories

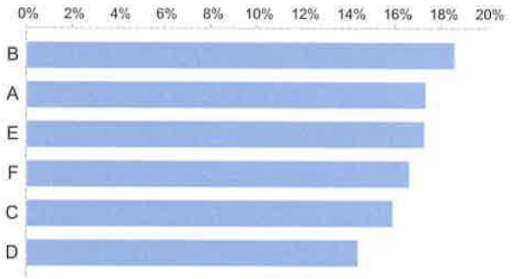
- Levels can be shares
- Emphasize rank order of levels
- Highlight one level relative to others

When Shape Doesn't Do What You'd Hoped



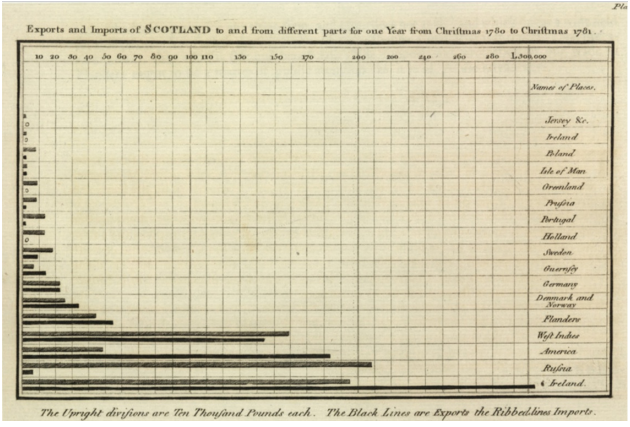
When Shape Doesn't Do What You'd Hoped





Few 6.11 and 6.12

The First Bar Chart



Playfair, William, 1786. *The Commercial and Political Atlas*

General Principles for Bar Charts

- Orientation
- Proximity
- Fills
- Borders
- Base value

Taken from Few Ch. 10, p. 210

Orientation & Proximity

Orientation: Bars horizontal or vertical?

- Horizontal better to fit in long lables
- Vertical better if axis is time

Orientation & Proximity

Orientation: Bars horizontal or vertical?

- Horizontal better to fit in long lables
- Vertical better if axis is time

Proximity – How close should the bars be?

- You want mostly bars, not mostly white space
- But not touching bars
- Why not touching bars?

Orientation & Proximity

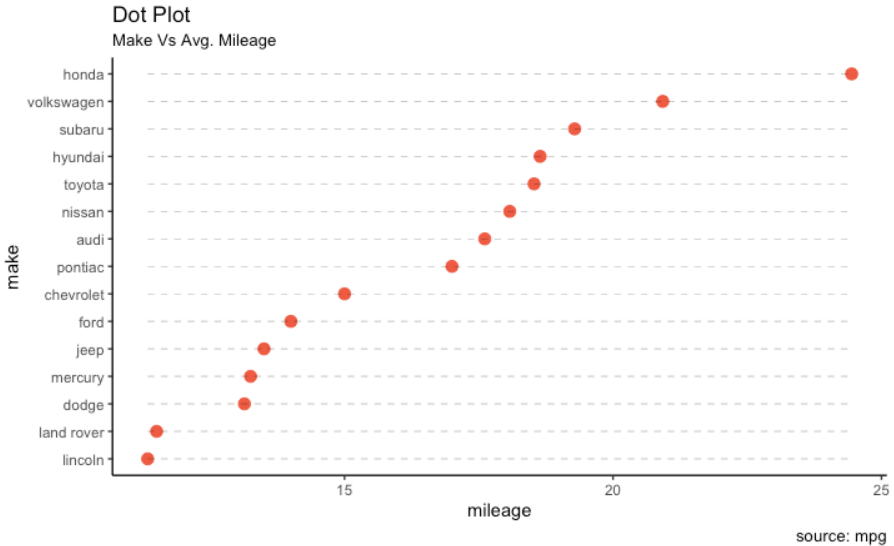
Orientation: Bars horizontal or vertical?

- Horizontal better to fit in long lables
- Vertical better if axis is time

Proximity – How close should the bars be?

- You want mostly bars, not mostly white space
- But not touching bars
- Why not touching bars?
- **Rank when you want to highlight ordering**

Ranked Almost-Bars



Fills

Fills

Do Not

- Use color as decoration
- Use hashed or lined fills

Do

- As much as possible, put legend directly on the graph
- Highlight with color

Fills

Fills

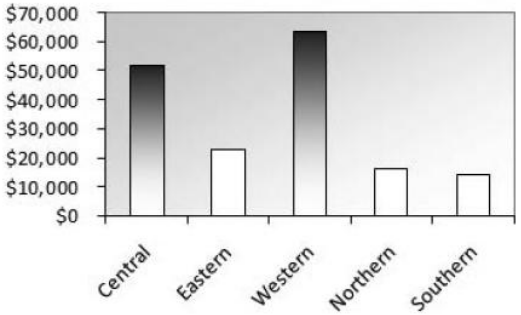
Do Not

- Use color as decoration
- Use hashed or lined fills

Do

- As much as possible, put legend directly on the graph
- Highlight with color

An Uninspiring Example



Fill source is Excel 2007

Borders and Base Value

Borders

- Use sparingly to highlight
- Colors are better for highlighting

Borders and Base Value

Borders

- Use sparingly to highlight
- Colors are better for highlighting

Base Value

Bars Must Start at Zero!

Borders and Base Value

Borders

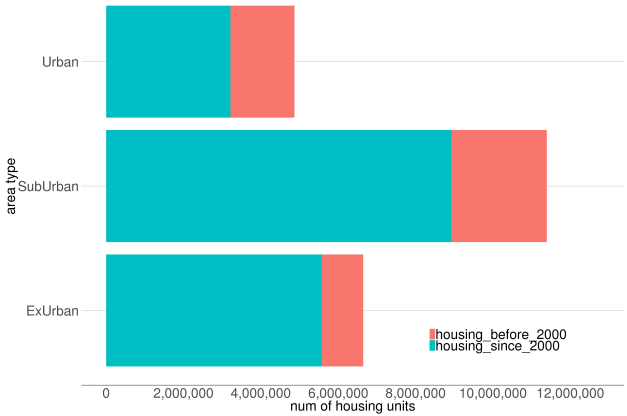
- Use sparingly to highlight
- Colors are better for highlighting

Base Value

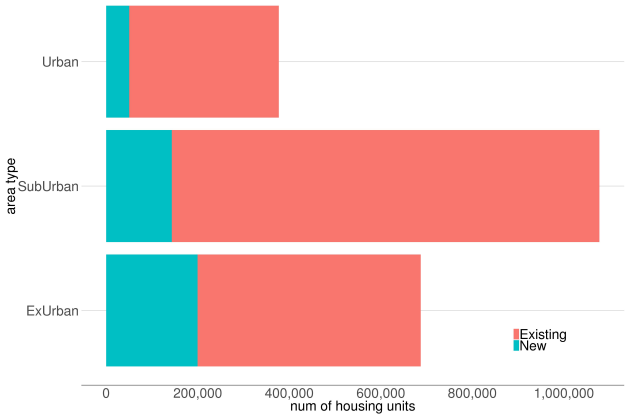
Bars Must Start at Zero!

Why?

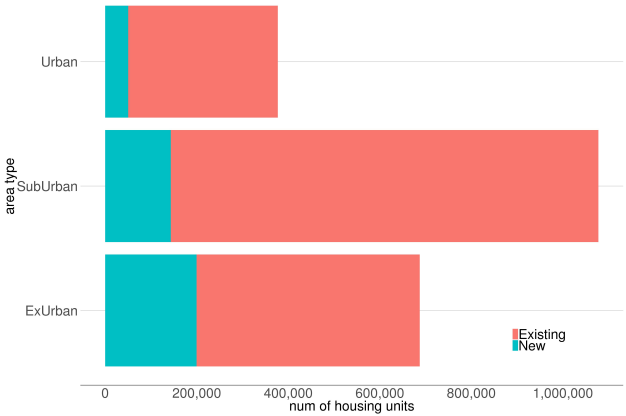
Going From a Bad Bar Chart to a Decent One



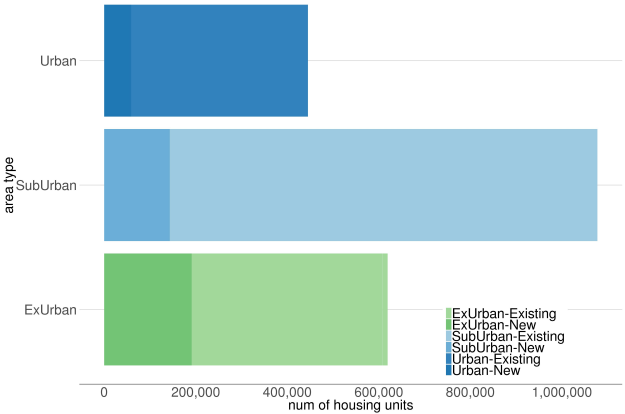
Going From a Bad Bar Chart to a Decent One



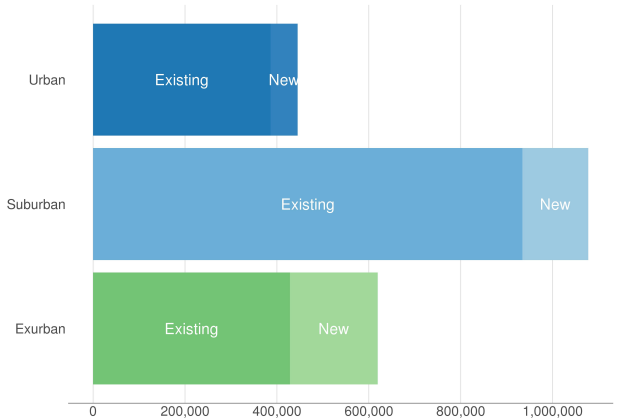
Going From a Bad Bar Chart to a Decent One



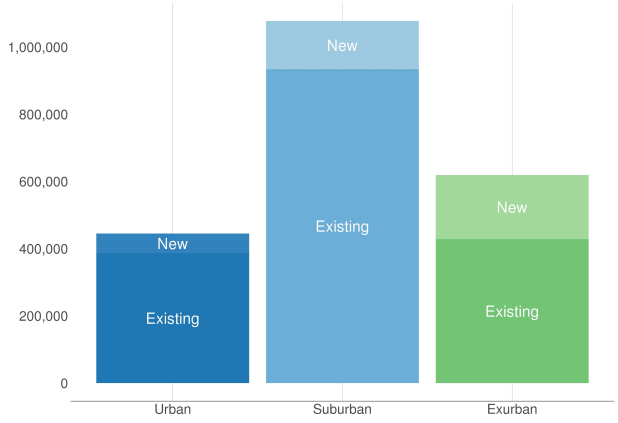
Going From a Bad Bar Chart to a Decent One



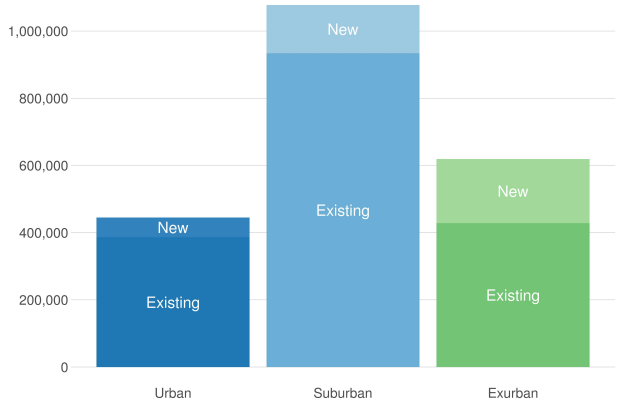
Going From a Bad Bar Chart to a Decent One



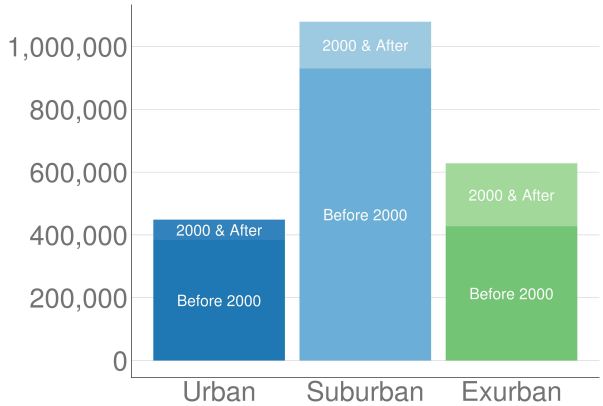
Going From a Bad Bar Chart to a Decent One



Going From a Bad Bar Chart to a Decent One



Going From a Bad Bar Chart to a Decent One



When the Number is Too Big for a Bar

People really have trouble with big numbers

- is $2 billion a large part of a $4 trillion budget?

When the Number is Too Big for a Bar

People really have trouble with big numbers

- is $2 billion a large part of a $4 trillion budget?

Microsoft plan for conveying big numbers, from *WSJ*

- attribute
- scaling factor
- reference

When the Number is Too Big for a Bar

People really have trouble with big numbers

- is $2 billion a large part of a $4 trillion budget?

Microsoft plan for conveying big numbers, from *WSJ*

- attribute
- scaling factor
- reference

"a conservation group that reclaimed about 100 million acres of land across the Earth. ... How big do you think that is?"

When the Number is Too Big for a Bar

People really have trouble with big numbers

- is $2 billion a large part of a $4 trillion budget?

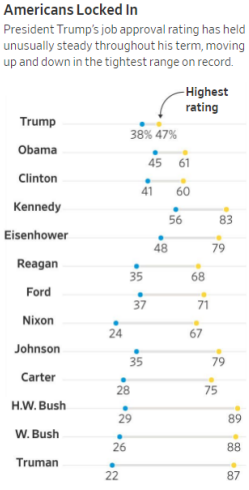
Microsoft plan for conveying big numbers, from *WSJ*

- attribute
- scaling factor
- reference

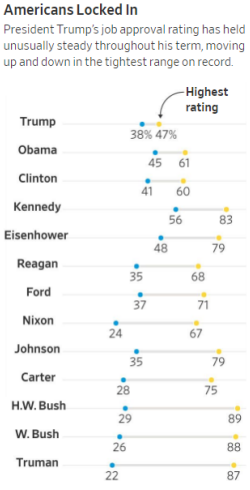
"a conservation group that reclaimed about 100 million acres of land across the Earth. ... How big do you think that is?"

About as big as (1.15x = scaling factor) California (reference)

The Power of Lollipops

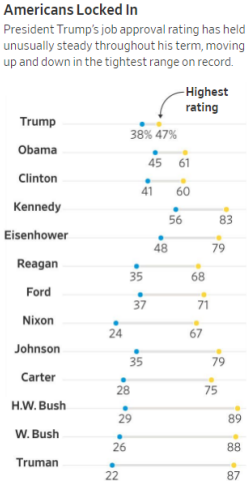


The Power of Lollipops



What is the point?

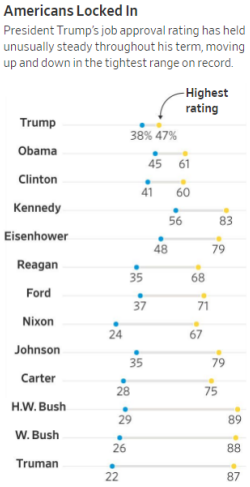
The Power of Lollipops



What is the point?

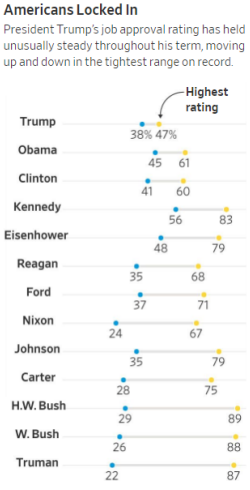
What is this doing that a bar cannot?

The Power of Lollipops



What is the point?
What is this doing that a bar cannot?
What info does this convey?

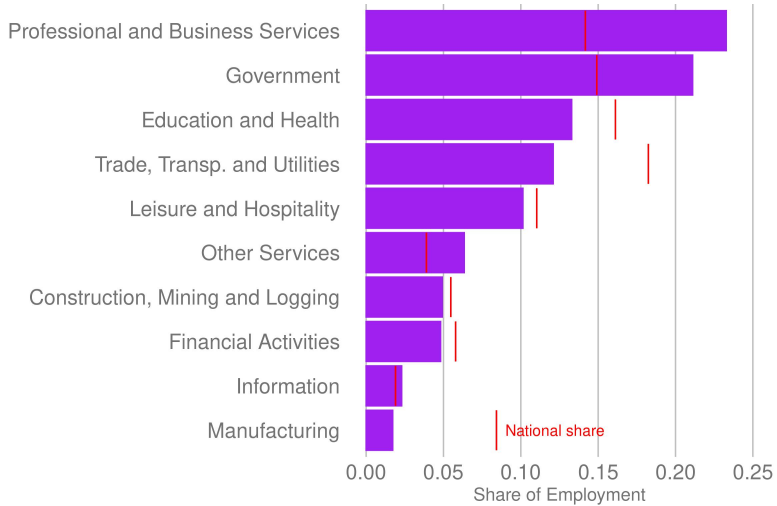
The Power of Lollipops



What is the point?
What is this doing that a bar cannot?
What info does this convey?

- max and min
- approximately the variance
- by administration
- so a trend in variance!
- note the point in the title

Showing More than One Number Per Bar



See you in class

R: Finally, Graphs!

Today

A. Graphing in R via `ggplot`

B. The parts of `ggplot`

C. Bars via `ggplot`

D. Titles and axis scaling

E. Factor re-ordering

F. Summary statistics

G. Date variables

A. Graphing in R

- there are 1000s of ways to graph in R
- we concentrate on `ggplot` – part of the `tidyverse`

A. Graphing in R

- there are 1000s of ways to graph in R
- we concentrate on `ggplot` – part of the `tidyverse`
- `ggplot` is the premier package for graphing in R
- There is a simple version of `ggplot` called `qplot`: we ignore it

A. Graphing in R

- there are 1000s of ways to graph in R
- we concentrate on `ggplot` – part of the `tidyverse`
- `ggplot` is the premier package for graphing in R
- There is a simple version of `ggplot` called `qplot`: we ignore it
- Developed in 2005 by Hadley Wickham
- In 2017, Wickham says "Ten years after ggplot2's release, Wickham wonders how much longer his program will dominate chart making in R. 'It really feels to me now like ggplot2 is ripe for disruption,' said Wickham. 'I'm surprised some young gun hasn't come along, and thought, 'Wow this is crap,' and done better. But so far, it hasn't really happened."'

Article link is here.

B. The Logic of `ggplot`

- Wickham was inspired by *The Grammar of Graphics*, by statistician Leland Wilkinson
- Builds on this logic to create plots

B. The Logic of `ggplot`

- Wickham was inspired by *The Grammar of Graphics*, by statistician Leland Wilkinson
- Builds on this logic to create plots
- Each plot has at least three elements
  1. **What** you want to graph: aesthetics, or `mapping = aes()`
  2. **How** you want to graph it: `geom_[something]`
  3. The overall **look**: zillions of commands, including axis modification

B. The Logic of `ggplot`

- Wickham was inspired by *The Grammar of Graphics*, by statistician Leland Wilkinson
- Builds on this logic to create plots
- Each plot has at least three elements
  1. **What** you want to graph: aesthetics, or `mapping = aes()`
  2. **How** you want to graph it: `geom_[something]`
  3. The overall **look**: zillions of commands, including axis modification
- Graphs can have > 1 geom – e.g., layer a scatter on top of a line

{With thanks to Andy Grogan-Kaylor}

Calling ggplot

```
ggplot() +
  geom_something(data = ,
                 mapping = aes(x = xvar, [y = yvar]))
```

Calling ggplot

```
ggplot() +
  geom_something(data = ,
                 mapping = aes(x = xvar, [y = yvar]))
```

This will pop up a graph in the plots window.

Making your graph an object

```
leahs.graph <- ggplot() +
  geom_something(data = ,
                 mapping = aes(x = xvar, [y = yvar]))
```

Will show nothing, but creates `leahs.graph` to which you can refer.

Making your graph an object

```
leahs.graph <- ggplot() +
  geom_something(data = ,
                 mapping = aes(x = xvar, [y = yvar]))
```

Will show nothing, but creates `leahs.graph` to which you can refer.
And

```
leahs.graph
```

Will pop up a graph.

You can add to an object

```
leahs.graph2 <- leahs.graph +
  geom_another(data = ,
                 mapping = aes(x = xvar, [y = yvar]))
```

I usually name the object and call the named object.

C.1. Bars

At its most basic

```
new.graph <- ggplot() +
  geom_col(data = [your data],
           mapping = aes(x = [categorical variable],
                         y = [value]))
```

C.1. Bars

At its most basic

```
new.graph <- ggplot() +
  geom_col(data = [your data],
           mapping = aes(x = [categorical variable],
                         y = [value]))
```

geom_col() plots the data you give – it doesn't calculate summary statistics from your data

`geom_bar()`: Bars Where `ggplot` Calculates For You

- instead of `geom_col()` you can use `geom_bar()`
- here R will add up the number of observations
- or take means
- but I find this syntax confusing and hard to check
- we will prepare data before `ggplot` and plot these data

C.2. Bar Chart Additions

Of course, there are many more things you can do

- make stacked bars: `position = "stack"`
- make grouped bars: `position = "dodge"`
- change the bar width
- change bar colors
- put labels on bars
- and still oodles more

D. Making Graphs Legible

```
new.graph <- ggplot() +
  geom_col(data = [your data],
           mapping = aes(x = [categorical variable],
                         y = [value])) +
  labs(title = "title here",
       x = "x label",
       y = "y label") +
  [things about scales] +
  theme([things you modify here])
```

D. Making Graphs Legible

```
new.graph <- ggplot() +
  geom_col(data = [your data],
           mapping = aes(x = [categorical variable],
                         y = [value])) +
  labs(title = "title here",
       x = "x label",
       y = "y label") +
  [things about scales] +
  theme([things you modify here])
```

+ 1000s of more options

E. Factor variables

- recall that R has a type of variable called a factor
- often created when a variable has a limited number of values
- useful to save memory space
- useful for making charts

E.1. Factor levels

- we particularly care about factor levels this class
- R orders bar charts by the order of the factor
- to change the order, change the order of the factor

E.2. Setting up a factor variable

```
states <- data.frame(state_abbrev = c("VA","DC","MD"),
                     state_fips = c(51,11,24),
                     av.feb.temp = c(50,47,45))
str(states)
```

```
## 'data.frame':    3 obs. of  3 variables:
##  $ state_abbrev: Factor w/ 3 levels "DC","MD","VA": 3 1 2
##  $ state_fips  : num  51 11 24
##  $ av.feb.temp : num  50 47 45
```

- state is a factor variable
- has three levels: DC, MD, VA
- in that order – R auto-alphabetizes
- suppose we prefer it in another order: VA, DC, MD

E.3. Re-ordering a factor

Change from [DC, MD, VA] to [VA, DC, MD]

```
levels(states$state_abbrev)
```

```
## [1] "DC" "MD" "VA"
```

```
states$state_abrev2 <- factor(states$state_abbrev,
                              levels = c("VA","DC","MD"))
levels(states$state_abrev2)
```

```
## [1] "VA" "DC" "MD"
```

E.3. Re-ordering a factor

Change from [DC, MD, VA] to [VA, DC, MD]

```
levels(states$state_abbrev)
```

```
## [1] "DC" "MD" "VA"
```

```
states$state_abrev2 <- factor(states$state_abbrev,
                              levels = c("VA","DC","MD"))
levels(states$state_abrev2)
```

```
## [1] "VA" "DC" "MD"
```

Remember: you need this to re-order bars.

F. Summary statistics are useful

- to check data
- to display data

F.1 Call dplyr package

```
library(dplyr)
```

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

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

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Part of the tidyverse. If not installed, you'll need to do so.

F.2. mutate()

- if you know Stata's egen, it's like that
- create a new variable that has the average temperature for all three states

```
library(dplyr)
```

```
states <- mutate(.data = states,
                 all.states.feb=mean(av.feb.temp,
                                     na.rm = TRUE))
states
```

```
##   state_abbrev state_fips av.feb.temp state_abrev2 all.states.feb
## 1           VA         51          50           VA       47.33333
## 2           DC         11          47           DC       47.33333
## 3           MD         24          45           MD       47.33333
```

F.2. mutate()

- if you know Stata's egen, it's like that
- create a new variable that has the average temperature for all three states

```
library(dplyr)
```

```
states <- mutate(.data = states,
                 all.states.feb=mean(av.feb.temp,
                                     na.rm = TRUE))
states
```

```
##   state_abbrev state_fips av.feb.temp state_abrev2 all.states.feb
## 1           VA         51          50           VA       47.33333
## 2           DC         11          47           DC       47.33333
## 3           MD         24          45           MD       47.33333
```

Why not just `av.temp <- mean(states$av.feb.temp, na.rm = TRUE)`?

F.3. More on `mutate()`

- it does many many other things as well
- you can use all kinds of functions in the second term
- and create more than one new variable
- can combine with `group_by()`

G. Date Variables: The Problem with Not Using Them

- if you have a time recorded as a character string: “2021-02-01”
- you can find the month and year
- but you can’t **correctly**
  - calculate differences between dates
  - use the variable in a chart

Date Variables: What They Do For You

- so software has something called a date variable
- usually seconds/minutes/hours/days since a point in time

Date Variables: What They Do For You

- so software has something called a date variable
- usually seconds/minutes/hours/days since a point in time
- if you have a date variable you can
  - calculate time or date differences
  - make charts with dates, used properly
  - and other valuable things
- today's tutorial walks through the creation of a date variable

Next Class

- I'm here till 5:20 for questions on anything
- Turn in PS 3
- Few, Chapter 6
- Chang, Chapter 6 (through 6.5)
- Linked Bloomberg article on quantities of land