Lecture 4: Histograms

February 10, 2020

Overview

Course Administration

Good, Bad and Ugly

Variations of Graphs, Few Ch. 9

What is a Histogram?

ggplot and Histograms

Course Administration

1. Return policy brief proposal comments at end of class
2. Reminder: Fully composed chart due **Feb. 24**
   - if there is something you want to do, but can't figure out how
   - write it in words accompanying the graph
3. Anything lingering?

General Policy Brief Proposal Feedback

Good work and interesting topics.

General Policy Brief Proposal Feedback

Good work and interesting topics.

- Remember: you need 5 to 8 graphics
- some basic descriptives often set the stage
- may be helpful to think about summary statistics before correlations
- aggregation does not mean merging together. it means going from one unit of observation to another
- with new data, good practice for you to match published summary stats
- as relevant, consider adding in decennial census/acs data to add demographics

General Policy Brief Proposal Feedback

Good work and interesting topics.

- Remember: you need 5 to 8 graphics
- some basic descriptives often set the stage
- may be helpful to think about summary statistics before correlations
- aggregation does not mean merging together. it means going from one unit of observation to another
- with new data, good practice for you to match published summary stats
- as relevant, consider adding in decennial census/acs data to add demographics
- expect to have problems

Next Week's Good Bad and Ugly

Find a histogram. Post by Wednesday noon (just do it this Wed. so you don't forget). You post the link on the google sheet. Earlier is ok.

- Boyd G.
- Didem B.
- Dallas C.

This Week's Good Bad and Ugly

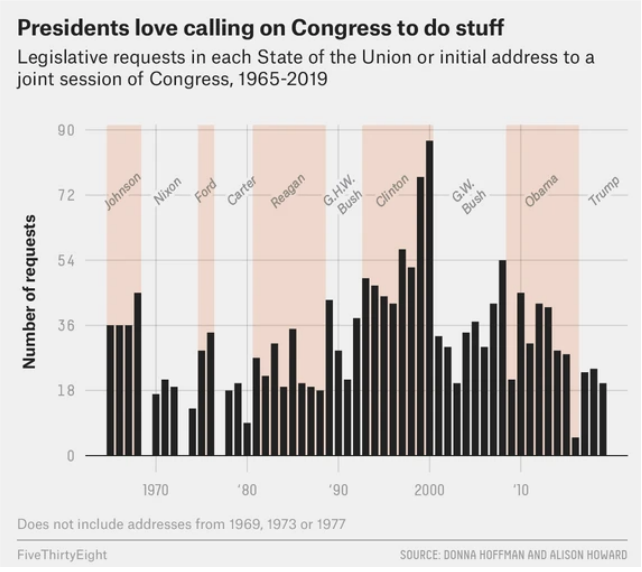
Finder	Commenter
Lauren G.	Boyd G.
Connor D.	Kim W.
Basia D.	Anna W.

Lauren's Example

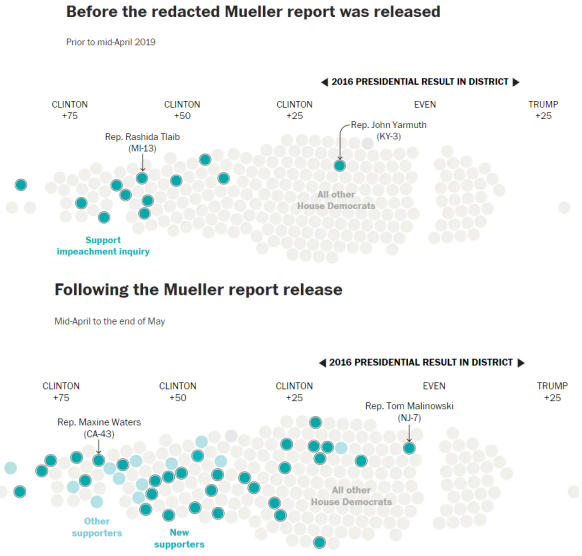
North America

Airline	Based in	Suspended	Dates of suspension
Air Canada	Canada	Flights to Beijing and Shanghai	Jan. 30 – Feb. 29
American Airlines	U.S.	All flights to China; and Hong Kong service from Dallas (from Feb. 1 to Feb. 21) and Los Angeles (Feb. 1 to March 27)	Jan. 31 – Mar. 27
Delta	U.S.	All flights to China	Feb. 2 – Apr. 30
United Airlines	U.S.	Service to Beijing, Shanghai and Chengdu; and Hong Kong service from Feb. 8 until Feb. 20	Feb. 5 – Mar. 28

Connor's Example



Basia's Example

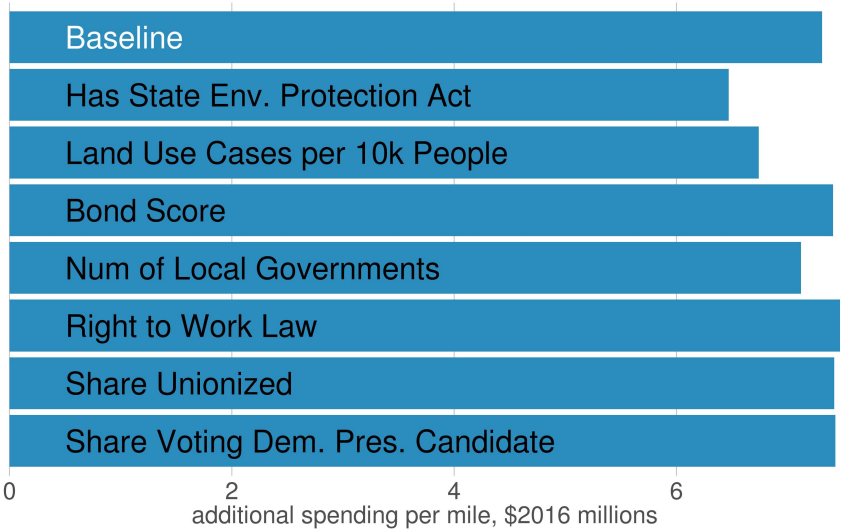


Which Graph for What Purpose?

Few: Three Basic Ways to Convey Information Graphically

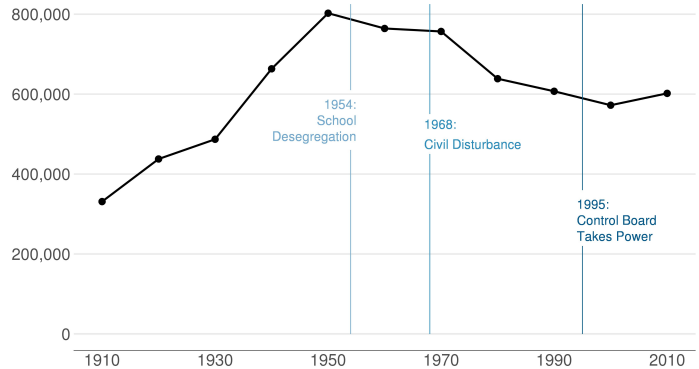
1. Bars
2. Lines
3. Boxes for distributions

Bars

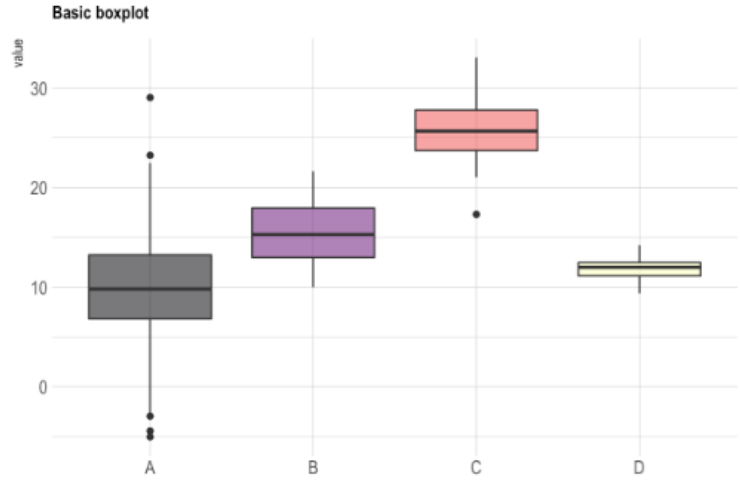

Baseline
Has State Env. Protection Act
Land Use Cases per 10k People
Bond Score
Num of Local Governments
Right to Work Law
Share Unionized
Share Voting Dem. Pres. Candidate
0
2
4
6
additional spending per mile, $2016 millions

Lines

Population Turns Up After 2000



Boxes



Source: https://www.r-graph-gallery.com/89-box-and-scatter-plot-with-ggplot2.html

Types of Relationships You May Want to Show, 1 of 2

Relationship	Use	Avoid
Nominal comparison		
Time Series		
Ranking		
Part-to-whole		

Types of Relationships You May Want to Show, 1 of 2

Relationship	Use	Avoid
Nominal comparison	Bars, Points sparingly	
Time Series		
Ranking		
Part-to-whole		

Types of Relationships You May Want to Show, 1 of 2

Relationship	Use	Avoid
Nominal comparison	Bars, Points sparingly	Bars starting above 0
Time Series		
Ranking		
Part-to-whole		

Types of Relationships You May Want to Show, 1 of 2

Relationship	Use	Avoid
Nominal comparison	Bars, Points sparingly	Bars starting above 0
Time Series	Lines	
Ranking		
Part-to-whole		

Types of Relationships You May Want to Show, 1 of 2

Relationship	Use	Avoid
Nominal comparison	Bars, Points sparingly	Bars starting above 0
Time Series	Lines	Bars falsely suggest independence
Ranking		
Part-to-whole		

Types of Relationships You May Want to Show, 1 of 2

Relationship	Use	Avoid
Nominal comparison	Bars, Points sparingly	Bars starting above 0
Time Series	Lines	Bars falsely suggest independence
Ranking	Bars or Dots	
Part-to-whole		

Types of Relationships You May Want to Show, 1 of 2

Relationship	Use	Avoid
Nominal comparison	Bars, Points sparingly	Bars starting above 0
Time Series	Lines	Bars falsely suggest independence
Ranking	Bars or Dots	Not lines!
Part-to-whole		

Types of Relationships You May Want to Show, 1 of 2

Relationship	Use	Avoid
Nominal comparison	Bars, Points sparingly	Bars starting above 0
Time Series	Lines	Bars falsely suggest independence
Ranking	Bars or Dots	Not lines!
Part-to-whole	Bars or stacked bars	

Types of Relationships You May Want to Show, 1 of 2

Relationship	Use	Avoid
Nominal comparison	Bars, Points sparingly	Bars starting above 0
Time Series	Lines	Bars falsely suggest independence
Ranking	Bars or Dots	Not lines!
Part-to-whole	Bars or stacked bars	No pies!

Types of Relationships You May Want to Show, 1 of 2

Relationship	Use	Avoid
Distribution		
Single		
Multiple		
Correlation		
Geospatial		

Types of Relationships You May Want to Show, 1 of 2

Relationship	Use	Avoid
Distribution		
Single	Histogram, dot plot, or density curve	
Multiple		
Correlation		
Geospatial		

Types of Relationships You May Want to Show, 1 of 2

Relationship	Use	Avoid
Distribution		
Single	Histogram, dot plot, or density curve	
Multiple	Bars or Dots	
Correlation		
Geospatial		

Types of Relationships You May Want to Show, 1 of 2

Relationship	Use	Avoid
Distribution		
Single	Histogram, dot plot, or density curve	
Multiple	Bars or Dots	Two histograms together is hard!
Correlation		
Geospatial		

Types of Relationships You May Want to Show, 1 of 2

Relationship	Use	Avoid
Distribution		
Single	Histogram, dot plot, or density curve	
Multiple	Bars or Dots	Two histograms together is hard!
Correlation	Points or paired bars	
Geospatial		

Types of Relationships You May Want to Show, 1 of 2

Relationship	Use	Avoid
Distribution		
Single	Histogram, dot plot, or density curve	
Multiple	Bars or Dots	Two histograms together is hard!
Correlation	Points or paired bars	Rarely lines
Geospatial		

Types of Relationships You May Want to Show, 1 of 2

Relationship	Use	Avoid
Distribution		
Single	Histogram, dot plot, or density curve	
Multiple	Bars or Dots	Two histograms together is hard!
Correlation	Points or paired bars	Rarely lines
Geospatial	Wait for maps!	

Histograms

Histogram Shows the Distribution of **One** Variable

- Take a variable
- Make bins by value
- Count the number of observations in each bin
- Plot bars with that number
- Walk through an example

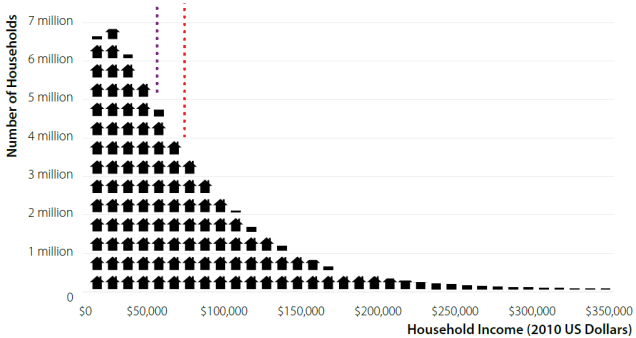
Key Features of Histograms

- A special case of a bar chart
- But! unlike a bar chart, histogram bars touch, to indicate continuity
- Which of Few's principles does this illustrate?

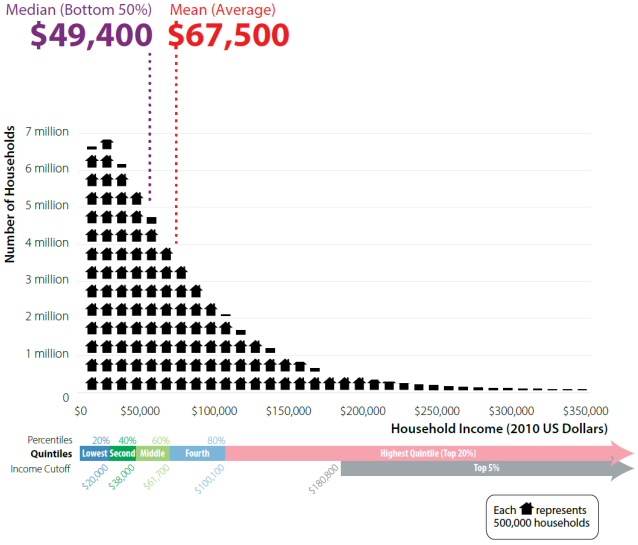
Some Examples

- Income distribution
- As a guide on a map
- Income distribution for DC MSA

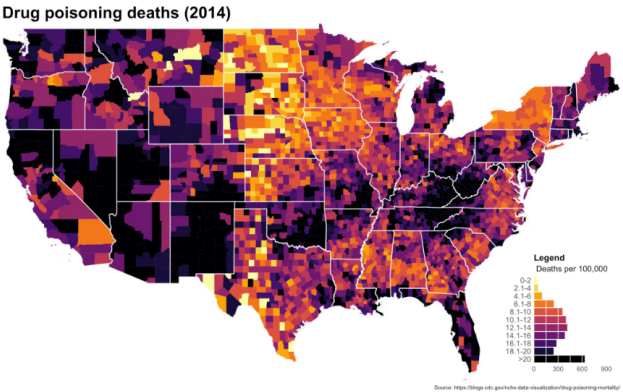
Mulbrandon's Income Histogram



Mulbrandon's Income Histogram



As a Map Legend

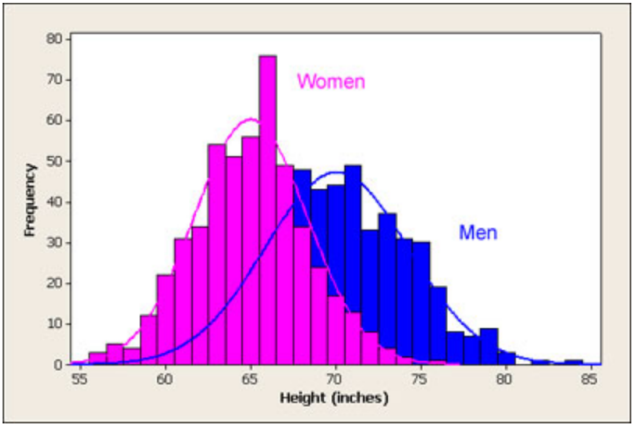


From https://mathewkiang.com/2017/01/16/using-histogram-legend-choropleths/

Density Curves: Smoothed Histograms

- Imagine many very thin bars
- This yields a curve
- Sometimes it is more helpful to draw the curve

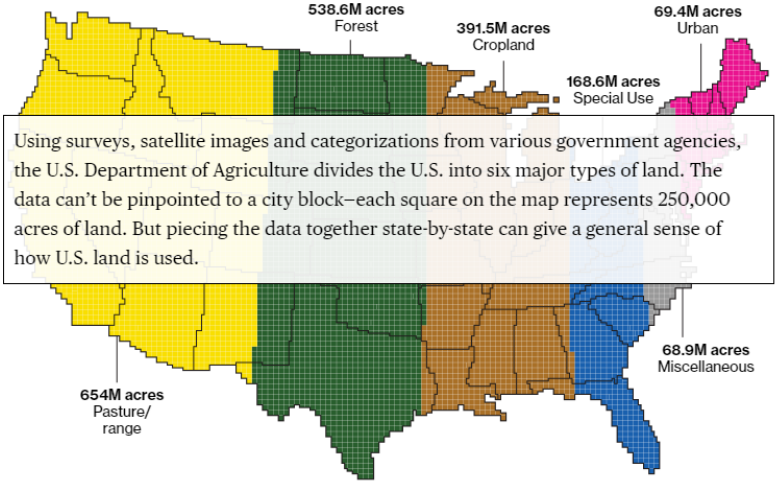
Height: Note the Curves



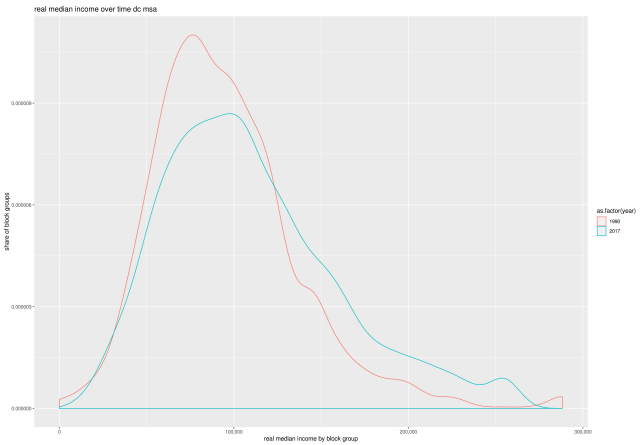
From http://www.usablestats.com/lessons/normal

Income Distribution in the DC Metro Area Over Time

Goal here is also histogram-like.



Income Distribution in the DC Metro Area Over Time



- need stronger lines
- note medians in each year
- get rid of grey background
- maybe add annotation to graph at right

R

Today

A. Heads-up: Bigger Data

B. `ifelse()` command

C. Histograms

D. Theme elements

A. Bigger Data

- You need to work with more data than you can see in a window
- Today's tutorial has techniques to do this
- Look to summary statistics

B. A Key Programming Command: `ifelse()`

```
df$var <- ifelse(test = [condition],
                 yes =  [do if condition true],
                 no =   [do if condition false])
```

B. An Example, 1 of 3

```
ex <- data.frame(building = c("A","B","C"),
                 yb = c("1983","1989","2005"))
ex
```

```
##   building   yb
## 1        A 1983
## 2        B 1989
## 3        C 2005
```

What if I want to know the century in which each building is built?

B. An Example, 2 of 3

```
ex$c <- ifelse(test = ex$yb < 2000,
                       yes = "20th",
                       no = "21st")
```

```
## Warning in Ops.factor(ex$yb, 2000): '<' not meaningful for factors
```

B. An Example, 2 of 3

```
ex$c <- ifelse(test = ex$yb < 2000,
                       yes = "20th",
                       no = "21st")
```

```
## Warning in Ops.factor(ex$yb, 2000): '<' not meaningful for factors
```

```
table(ex$c)
```

```
## < table of extent 0 >
```

B. An Example, 2 of 3

```
ex$c <- ifelse(test = as.numeric(as.character(ex$yb)) < 2000,
                      yes = "20th",
                      no = "21st")
```

B. An Example, 2 of 3

```
ex$c <- ifelse(test = as.numeric(as.character(ex$yb)) < 2000,
                      yes = "20th",
                      no = "21st")
```

```
table(ex$c)
```

```
##
## 20th 21st
##    2    1
```

B. An Example, 2 of 3

```
ex$c <- ifelse(test = as.numeric(as.character(ex$yb)) < 2000,
                      yes = "20th",
                      no = "21st")
```

```
table(ex$c)
```

```
##
## 20th 21st
##    2    1
```

What could go wrong with programming like this?

B. Some rules of thumb for `ifelse()`

- check your output!

B. Some rules of thumb for `ifelse()`

- check your output!
- a test can include multiple conditions
- good idea to define all cases – don't let a case be the residual

B. Some rules of thumb for `ifelse()`

- check your output!
- a test can include multiple conditions
- good idea to define all cases – don't let a case be the residual
- you can nest `ifelse()` commands:

```
ex$ybn <- as.numeric(as.character(ex$yb))
summary(ex$ybn)
ex$c <- ifelse(test = ex$ybn >= 1900 & ex$ybn < 2000,
               yes = "20th",
               no = ifelse(test = ex$ybn >= 2000 & ex$ybn < 2100)
                           yes = "21st"
                           no = "trouble"))
```

C. Discrete Histograms

For discrete distributions, use

```
geom_histogram(data = [dataframe],
               mapping = aes(x = [variable]))
```

C. Discrete Histograms

For discrete distributions, use

```
geom_histogram(data = [dataframe],
               mapping = aes(x = [variable]))
```

Many options include

- fill: inside `aes`, `fill = [variable]`
- bin width: `bin_width = [unit span]`,
- by groups: inside `aes`, `color = [grouping variable]`
- facet to make small multiples: `+ facet_wrap([grouping variable])`

C. Approximating Continuous Distributions

For almost-continuous bins, use

```
geom_freqpoly()
```

For much more smoothing, use

```
geom_density()
```

D. Themes

- theme is the look and feel of the graph
- themes have **zillions** of elements
- see here for full list
- basic idea is that every part of the plot has a name
- you reference the name and tell R what to do with it

D. Changing Elements of a Theme

```
p1 <- ggplot() +
  geom_histogram(data = df,
                 mapping = aes(x = xvar)) +
  theme(
    panel.background = element_blank()
  )
```

- you can set theme elements to
- `element_line()`
- `element_text()`
- `element_rect()`

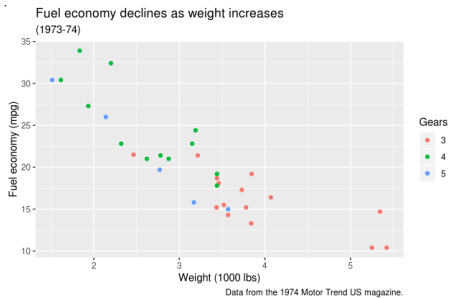
D. Pre-set Themes

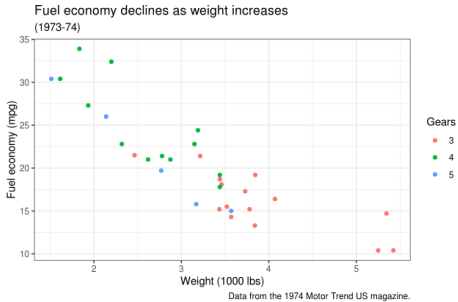
- you can get a head start with pre-set themes
- just add them to your `ggplot()` command:

```
p1 <- ggplot() +
  geom_histogram(data = df,
                 mapping = aes(x = xvar)) +
  theme_bw()
```

- see the list of 10 here
- you can also use the pre-set themes and then modify

D. Pre-set Theme Example





Next Class

No class next Monday – enjoy Presidents Day. On Feb. 24

- Turn in PS 4
- Turn in fully composed chart assignment to google folder
- Monmonier, *How to Lie with Maps*, Chapters 1 and 2
- Look at linked dot density map from *Post*