Lecture 4: Histograms

February 8, 2021

Overview

Course Administration

Good, Bad and Ugly

Variations of Graphs, Few Ch. 9

What is a Histogram?

R: ggplot and Histograms

Course Administration

1. Policy brief proposal comments online in your folder

Course Administration

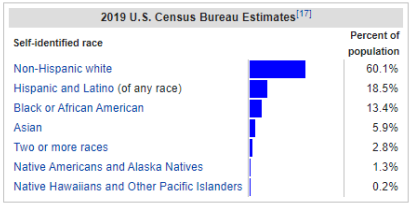
1. Policy brief proposal comments online in your folder
2. Reading quiz To Q1 pictures

Course Administration

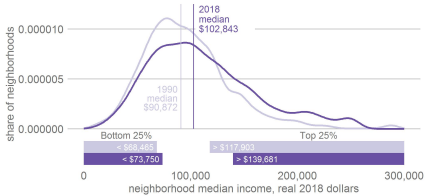
1. Policy brief proposal comments online in your folder
2. Reading quiz To Q1 pictures
3. Reminder: Fully composed chart due **Feb. 22**
   - if there is something you want to do, but can't figure out how
   - write it in words accompanying the graph
4. Anything lingering?

Charts for Q1

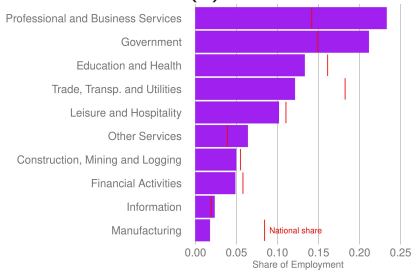
(a), from Wikipedia



(b), *State of the Capital Region*, 2020



(c)



State of the Capital Region, 2021

return to admin

Tutorial Feedback

- write up your answers: one document that clearly lists question and answer
- don't make me (or anyone) have to read code
- give evidence that you have worked through the tutorial
  - R code
  - R output
- R Markdown instructions posted if curious – resources tab

General Policy Brief Proposal Feedback

Good work and interesting topics. Feedback in your tutorial folder.

General Policy Brief Proposal Feedback

Good work and interesting topics. Feedback in your tutorial folder.

Successful proposals

- clearly set out the ≥ 2 data sources you're using
- how you're planning on aggregating data
- gave a sense of having some thoughts about the graphics you'd like to do or the points you'd like to make
- for this class, aggregation does not mean merging together. it means going from one unit of observation to another

Looking forward to the final product

- Final product needs 5 to 8 graphics
- some basic descriptives often set the stage
- may be helpful to think about summary statistics before correlations
- with new data, good practice for you to match published summary stats
- as relevant, consider adding in decennial census/acs data to add demographics

Looking forward to the final product

- Final product needs 5 to 8 graphics
- some basic descriptives often set the stage
- may be helpful to think about summary statistics before correlations
- 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

Looking forward to the final product

- Final product needs 5 to 8 graphics
- some basic descriptives often set the stage
- may be helpful to think about summary statistics before correlations
- 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 deadline: Lecture 5, one fully composed chart

Next Week's Good Bad and Ugly

Find a histogram. Post by Wednesday noon. Post the link on the google sheet.

Finder	Commenter
Winnie W.	Eleanor T.

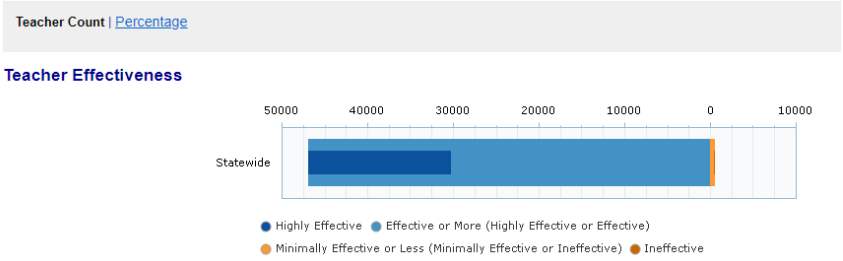
This Week's Good Bad and Ugly

Finder	Commenter
Trinday	Sarah
Eleanor	Preston

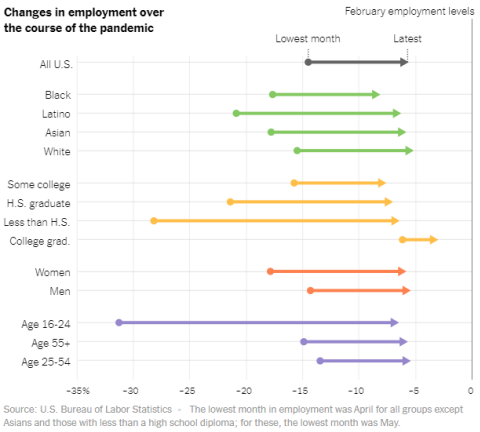
Trinday's Graph

Data from Spotify suggest that listeners are gloomiest in February

"Educator Effectiveness" from MI School Data



Eleanor's Graph



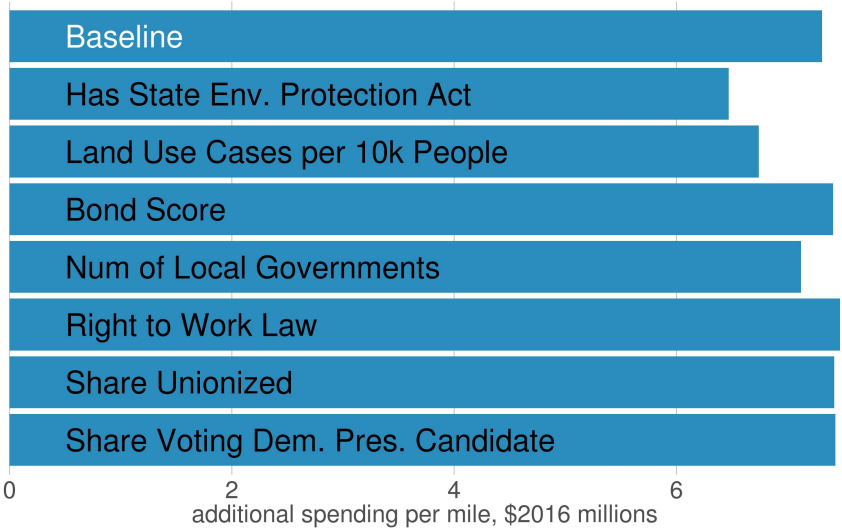
Opinion: The Year in Charts, *New York Times*

Which Graph for What Purpose?

Few: Three Basic Ways to Convey Information Graphically

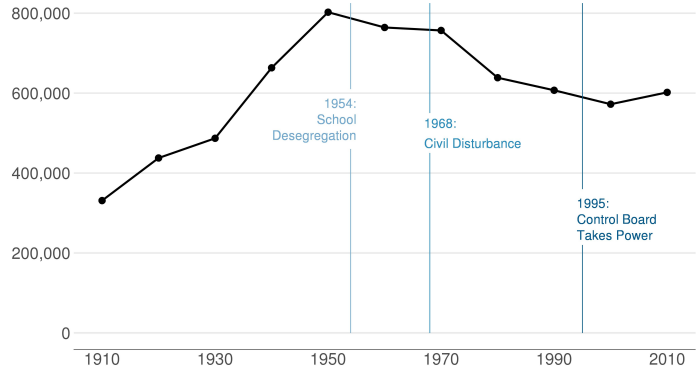
1. Bars
2. Lines
3. Boxes for distributions

Bars

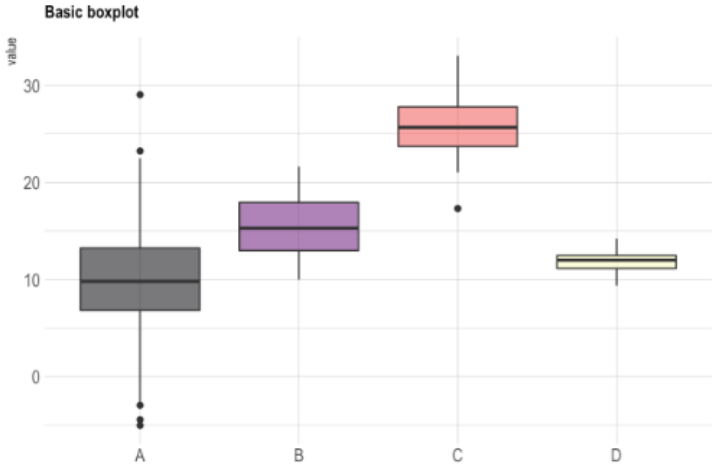

Baseline
Has State Env. Protection Act
Land Use Cases per 10k People
Bond Score
Num of Local Governments
Right to Work Law
Share Unionized
Share Voting Dem. Pres. Candidate
0
2
4
6
additional spending per mile, $2016 millions

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

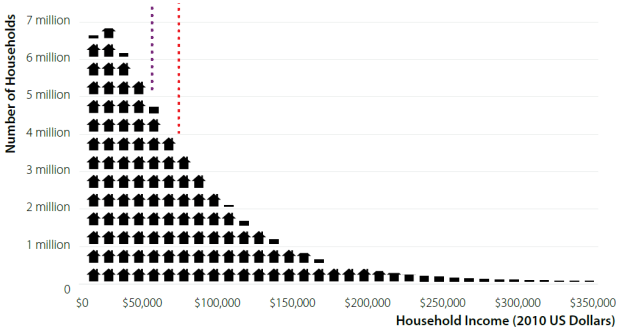
Key Features of Histograms

- 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?

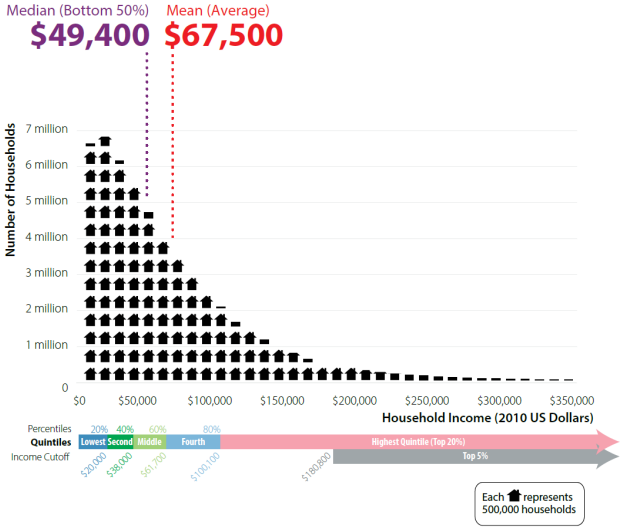
Some Examples

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

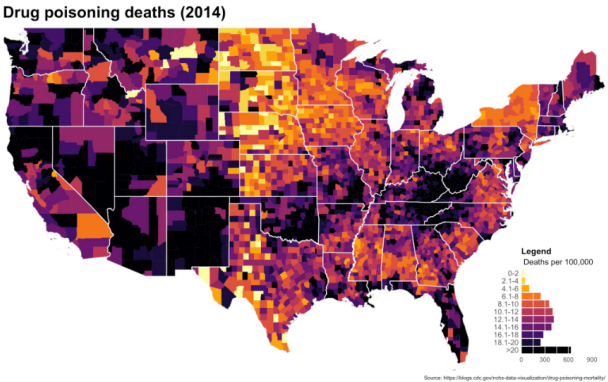
Mulbrandon's Income Histogram



Mulbrandon's Income Histogram



As a Map Legend

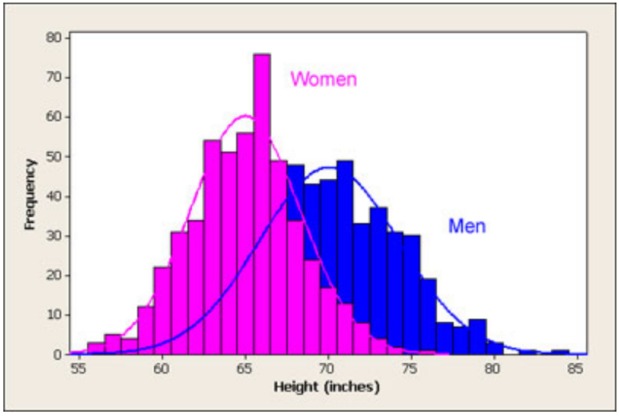


From https://mathewkiang.com/2017/01/16/using-histogram-legend-choropleths/

Density Curves: Smoothed Histograms

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

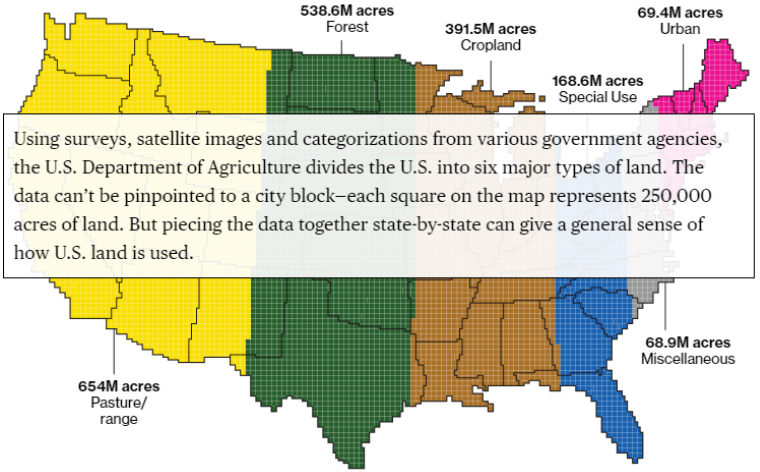
Height: Note the Curves



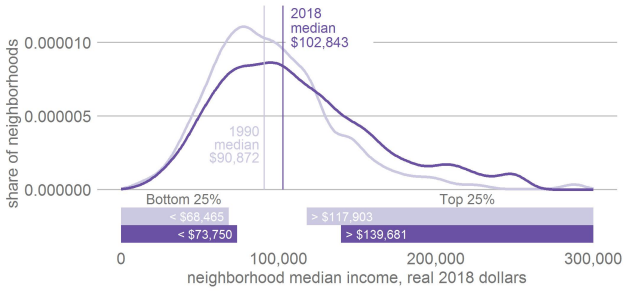
From http://www.usablestats.com/lessons/normal

Income Distribution in the DC Metro Area Over Time

Goal here is also histogram-like.



Income Distribution in the DC Metro Area Over Time



- was never satisfied with y axis
- light purple probably too light
- goal was to show 25th and 75th percentiles
- and change therein

R

To end of lecture

Today

A. Heads-up: Bigger Data

B. If-else recap

C. Histograms

D. Results by group: groupings and facets

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

A. Looking at crashes

```
dim(crash)
```

```
## [1] 59777    44
```

```
table(crash$Light)
```

```
##
## DARK -- UNKNOWN LIGHTING           DARK LIGHTS ON           DARK NO LIGI
##                      660                    13971                       2
##                     DAWN                 DAYLIGHT                      DU
##                     1239                    39305                       1
##                      N/A                    OTHER                   UNKN(
##                      497                      143                       4
```

- look at the total size of the dataset

A. A Legible Version

```
## # A tibble: 9 x 2
##   Light                    light_type
##   <fct>                         <int>
## 1 DARK -- UNKNOWN LIGHTING        660
## 2 DARK LIGHTS ON                13971
## 3 DARK NO LIGHTS                 2158
## 4 DAWN                           1239
## 5 DAYLIGHT                      39305
## 6 DUSK                           1393
## 7 N/A                             497
## 8 OTHER                           143
## 9 UNKNOWN                         411
```

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

```
df$var <- ifelse(test = [condition with ==],
                 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, 3 of 3

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

B. An Example, 3 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, 3 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. Histograms

We will use three new geoms this lecture

- `geom_histogram()`
- `geom_density()`
- `geom_freqpoly()`

C.1. How to create a histogram

Use

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

- only need to list one variable
- histograms are univariate graphics
- `geom_histogram()` is best for a distribution with limited values

C.1. How to create a histogram

Use

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

- only need to list one variable
- histograms are univariate graphics
- `geom_histogram()` is best for a distribution with limited values
- but not a categorical distribution, which should be a bar

C.2. Histogram options

- fill overall: outside `aes`, `fill = [color]`
- fill by group: inside `aes`, `fill = [variable]`
- bin width: `bin_width = [unit span]`,
- by groups: inside `aes`, `color = [grouping variable]`

C.3. Approximating Continuous Distributions

For almost-continuous bins, use

```
geom_freqpoly()
```

For much more smoothing, use

```
geom_density()
```

C.4. Example

- take crash-level data from last class
- use `group_by()` and `summarize()` to make daily data
- count number of crashes by day

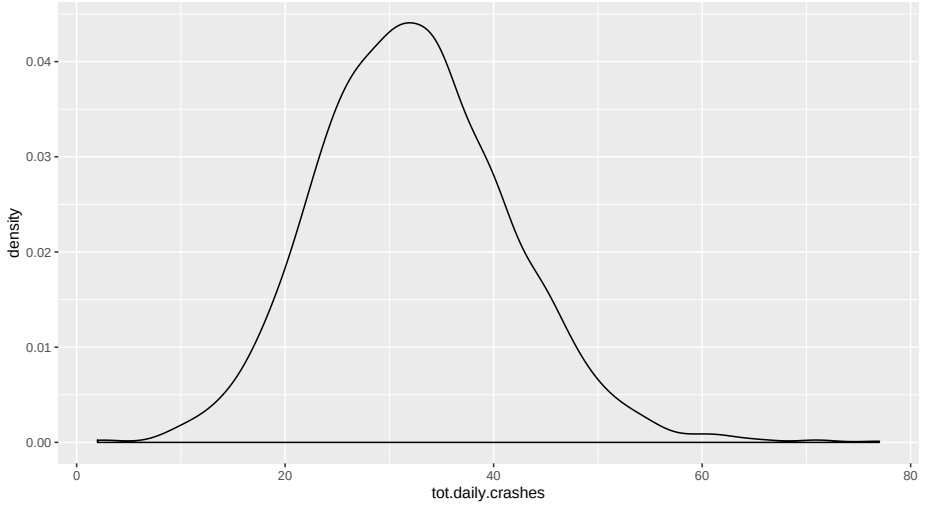
```
# add up total number of crashes by date
crash2 <- group_by(.data = crash, date2)
crash2 <- summarize(.data = crash2, tot.daily.crashes = n())
table(crash2$tot.daily.crashes)
```

```
##
##  2  3  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
##  1  1  1  2  4  5  3  8  8 11 12 20 19 33 26 44 43 57 61 65 76 73 76 74
## 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55
## 80 76 85 89 66 59 61 55 59 48 38 27 37 39 18 25 18 11 14  8  8  6  6  5
## 59 60 61 62 63 65 66 71 77
##  1  2  2  2  1  1  1  2  1
```

Plot these data

```
alld <- ggplot() +
  geom_density(data = crash2,
               mapping = aes(x = tot.daily.crashes))
```

Plot these data



D. Results by Group

```
# find the day of the week
crash2$day.of.week <- weekdays(x = crash2$date2)

# check
table(crash2$day.of.week)
```

```
##
##    Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
##       264       264       264       264       265       264       264
```

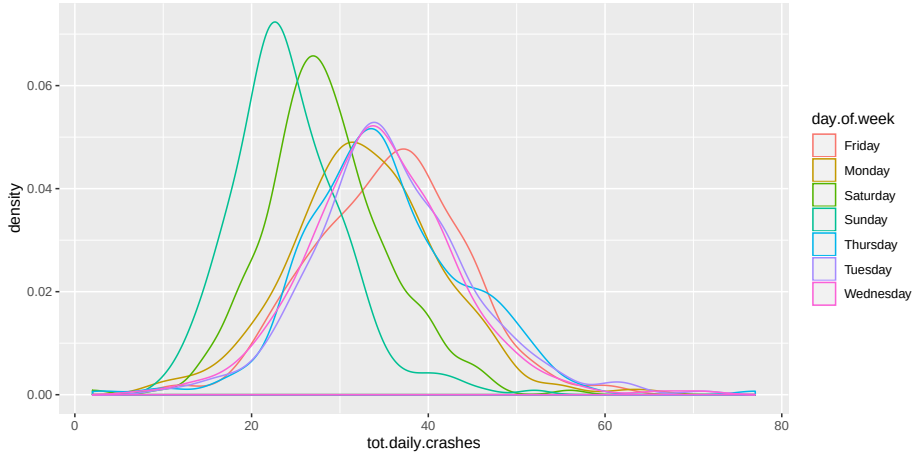
- you need a variable that indicates a group
- then plot distribution by group
- we'll use distribution of traffic accidents (x variable)
- by weekday (grouping variable)

By day of the week

```
wd <- ggplot() +
  geom_density(data = crash2,
               mapping = aes(x = tot.daily.crashes,
                             color = day.of.week))
```

By day of the week

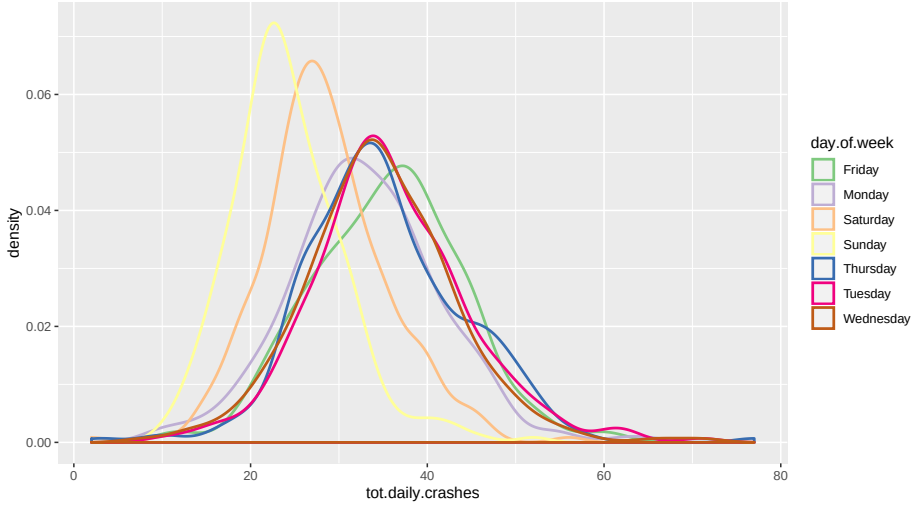
```
wd
```



By day of the week, better colors and thicker lines

```
day.colors <- c('#7fc97f','#beaed4','#fdc086','#ffff99',
                '#386cb0','#f0027f','#bf5b17')
wd <- ggplot() +
  geom_density(data = crash2,
               mapping = aes(x = tot.daily.crashes,
                             color = day.of.week),
               size = 0.9) +
  scale_color_manual(values = day.colors)
```

By day of the week, better colors and thicker lines

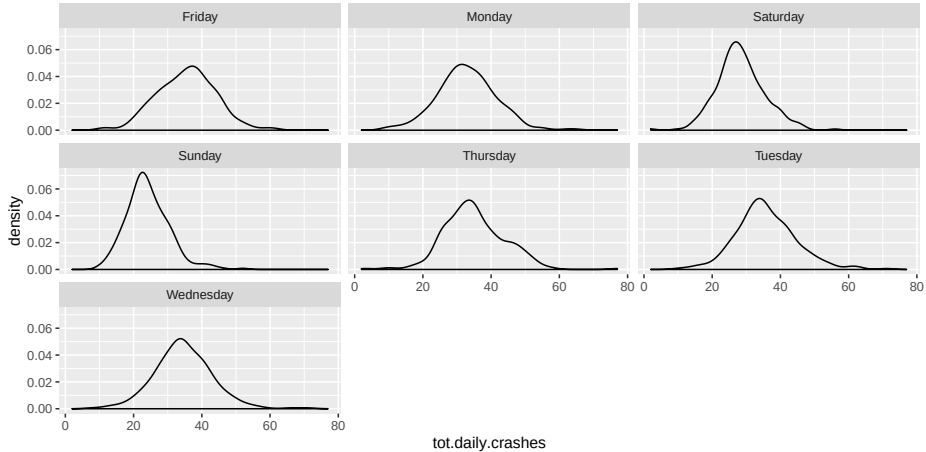


By day of the week, facets

```
wd <- ggplot() +
  geom_density(data = crash2,
               mapping = aes(x = tot.daily.crashes,
               group = day.of.week)) +
  facet_wrap(~day.of.week)
```

By day of the week, facets

```
wd
```



Next Class

No class next Monday – enjoy Presidents Day.

On Feb. 22

- Turn in Tutorial 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*