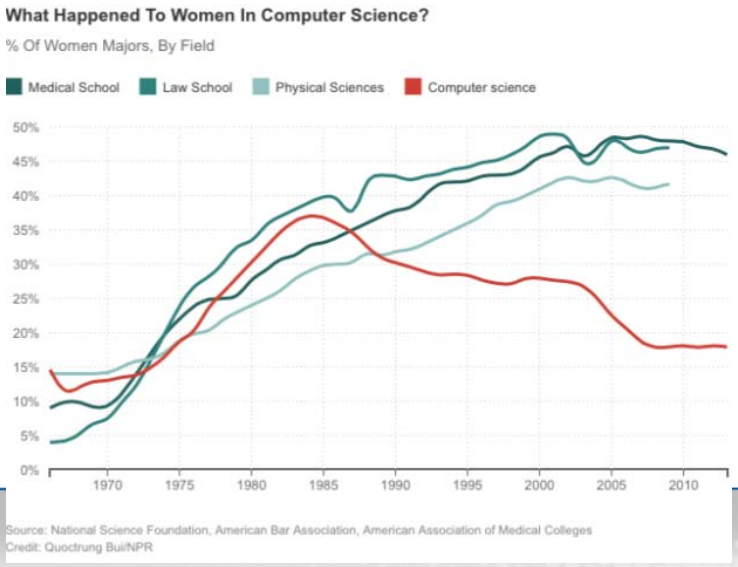
Data Visualization

ggplot2

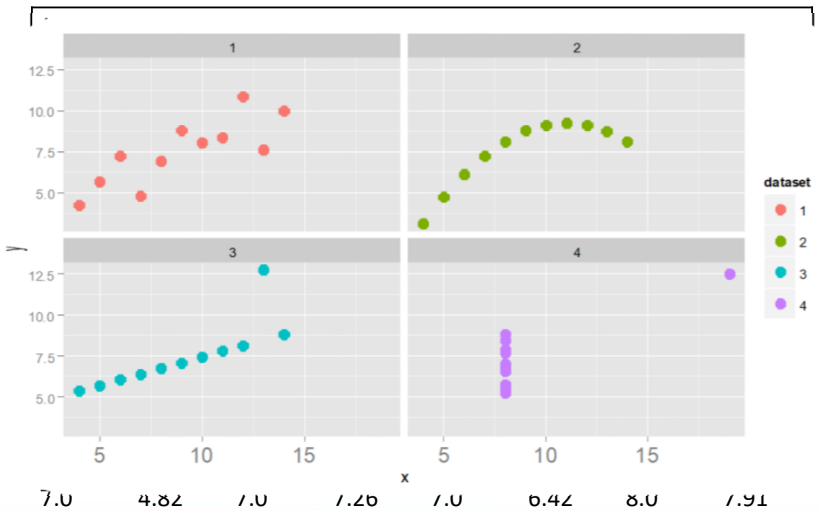
UNIVERSITY OF DELAWARE

Data Visualization
Why and How?

Why Visualization? Visualization Makes A Lot of Information Easily Digestable



Why Visualization? Accessiblity

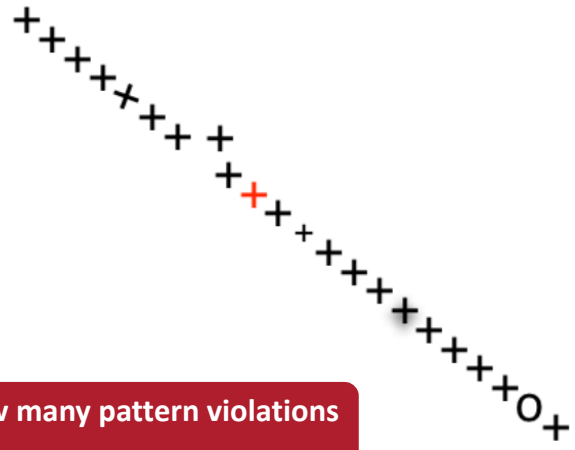


7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Note: Each x has the same mean and each y has the same mean

Why Visualization?

- Leverages brain and eyes ability to find patterns and pattern violations:
  - Trends
  - Gaps
  - Outliers



Question: How many pattern violations do you see?

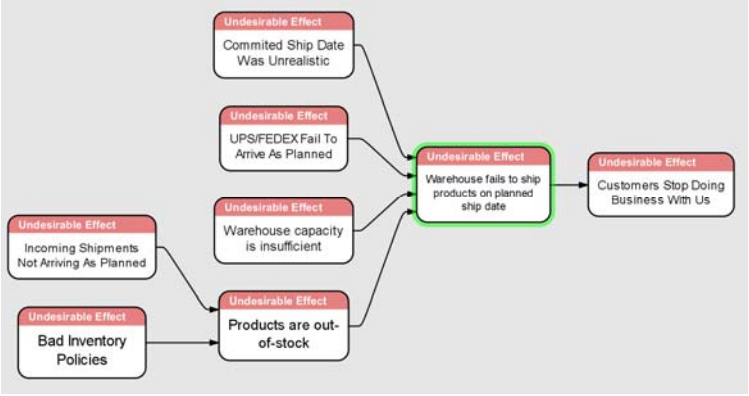
Four Steps of Data Visualization

1) Purpose – why are we using a visual
2) Content – what information should be shown
3) Structure – how do we visualize the content
4) Formatting – making it more informative, persuasive, and/or aesthetic.

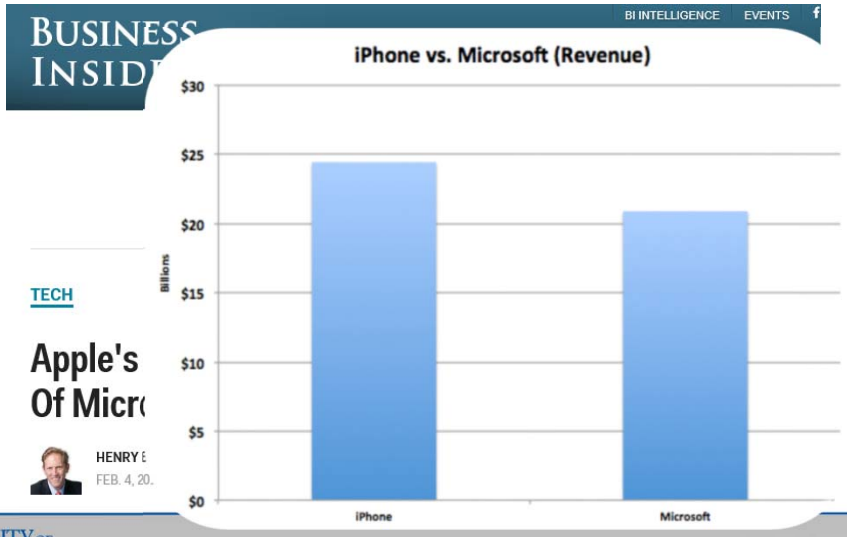
Step 1: Purpose
Always remember your purpose

- Which map is better for a 2-week kayaking trip along the coast of Greenland?

Five Why's & Current Reality Tree
Help Discover The Right Purpose



Step 2: Content



UNIVERSITY OF DELAWARE

Step 2: Content

1) What data matters?
2) What relationships matter?
3) Informed by purpose!
4) What's excluded is as important as what's included.

UNIVERSITY OF DELAWARE

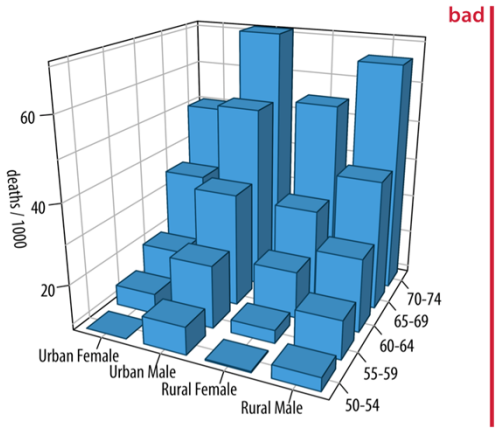
http://demographics.coopercenter.org/DotMap/index.html

Filtering Data for Better Clarity

Step 3: Structure

1) How do we best reveal the most important relationships?
2) Choose meaningful layout and axes!
3) Use both axes! (Both, not three…)
4) Informed by purpose and content

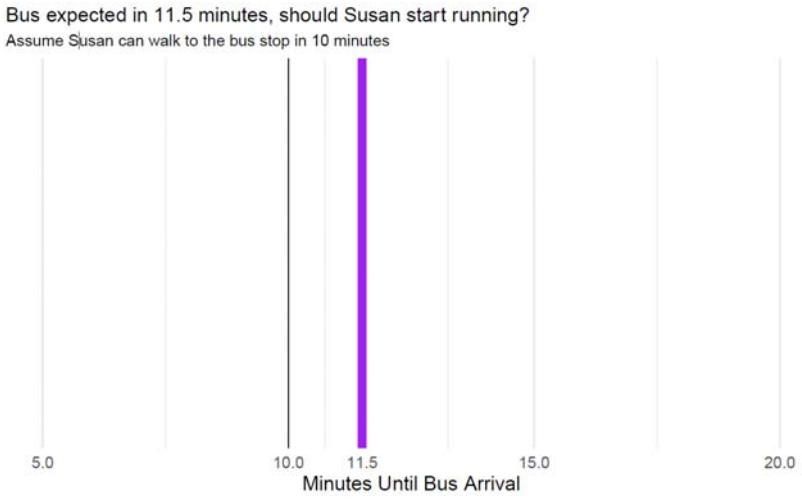
Structure: 3D is Bad



Author is trying to show mortality rate is higher for males than females. Is this easy to see?

Source: https://serialmentor.com/dataviz/no-3d.html

Structure: Visualizing Uncertainty



Structure: Visualizing Uncertainty

Structure: Visual Encodings

Properties and Best Uses of Visual Encodings

Example	Encoding	Ordered	Useful values	Quantitative	Ordinal	Categorical	Relational
	position, placement	**yes**	**infinite**	Good	Good	Good	Good
1, 2, 3; A, B, C	text labels	**optional** (alphabetical or numbered)	**infinite**	Good	Good	Good	Good
	length	**yes**	**many**	Good	Good		
	size, area	**yes**	**many**	Good	Good		
	angle	**yes**	medium/few	Good	Good		
	pattern density	**yes**	few	Good	Good		
	weight, boldness	**yes**	few		Good		
	saturation, brightness	**yes**	few		Good		
	color	no	few (< 20)			Good	
	shape, icon	no	medium			Good	
	pattern texture	no	medium			Good	
	enclosure, connection	no	**infinite**			Good	Good
	line pattern	no	few				Good
	line endings	no	few				Good
	line weight	**yes**	few		Good		

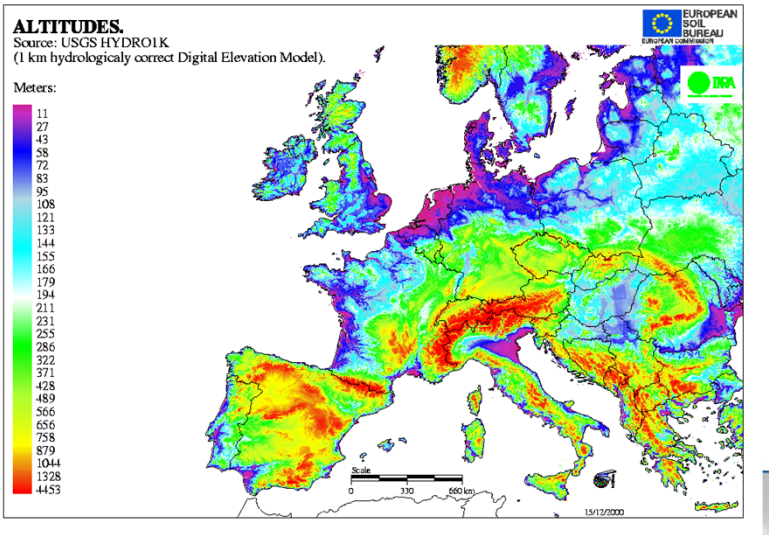
Quantitative = numeric

Ordinal = ranked or time

Categorical = groups

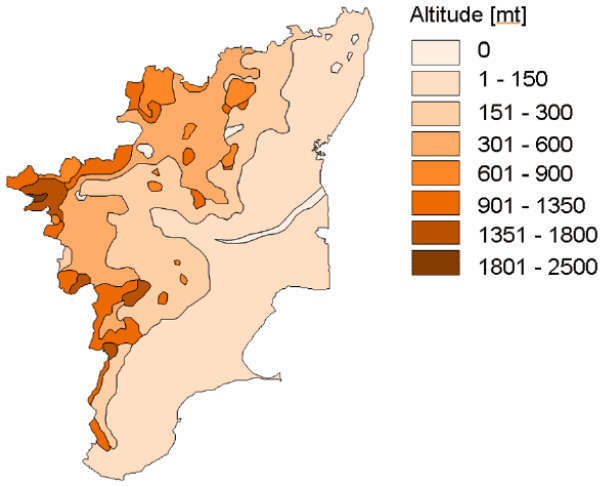
Relational = hierarchy or influence

Formatting: Color is not good for numeric or ordered data



UNIVERSITY OF DELAWARE

Formatting: Saturation or Hue is Better for Ordered Data



UNIVERSITY OF DELAWARE

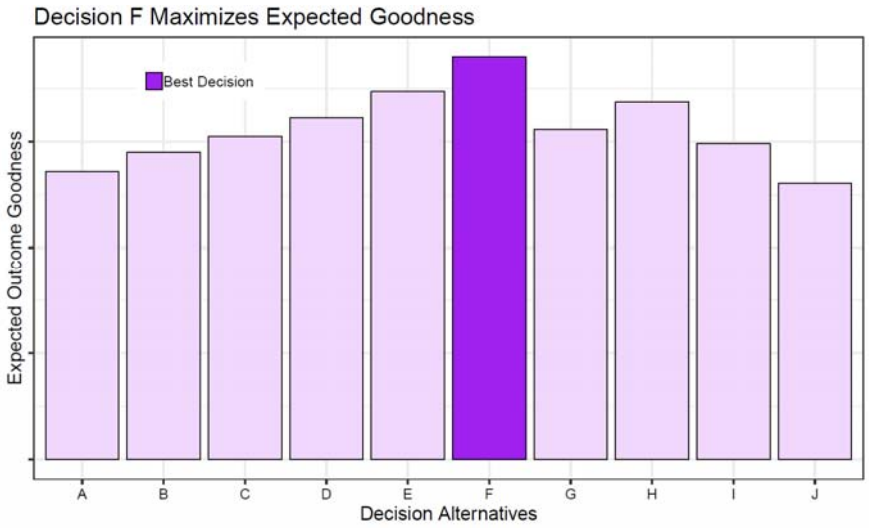
Color is better for categorical data

Step 4: Formatting

Make it beautiful

1) Highlight the most meaningful data
2) Use ink and color in proportion to importance

What action am I trying to persuade you to do?



Data Visualization

- Mastering the grammar of graphics
  - All about mapping
    - Use `aes( )` to map data to visual elements

Complete the template below to build a graph.

```
ggplot(data = <DATA>) +                                   ] required
<GEOM_FUNCTION>(mapping = aes(<MAPPINGS>),
 stat = <STAT>, position = <POSITION>) +                  ] Not required, sensible defaults supplied
<COORDINATE_FUNCTION> +
<FACET_FUNCTION> +
<SCALE_FUNCTION> +
<THEME_FUNCTION>
```

Example:

```
ggplot(data = mpg) +
   geom_point(mapping = aes(x = cty, y = hwy))
```

Data Visualization

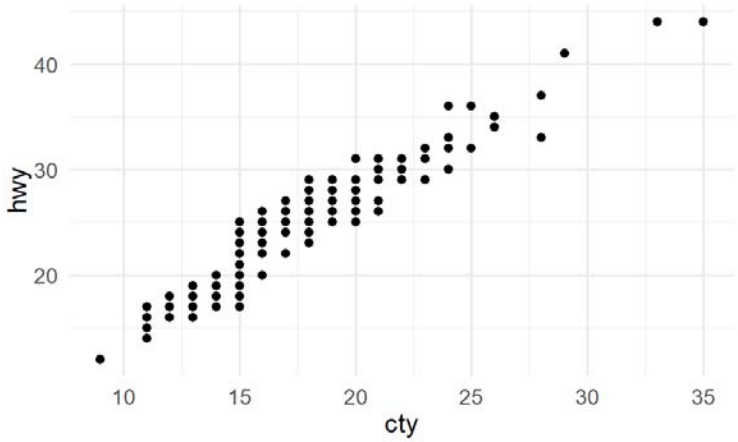
- Mastering the grammar of graphics
  - All about mapping
    - Use `aes( )` to map data to visual elements

Complete the template below to build a graph.

```
ggplot (data = <DATA>) +                                  ] required
<GEOM_FUNCTION> (mapping = aes( <MAPPINGS> ),
 stat = <STAT> , position = <POSITION> ) +                ] Not required, sensible defaults supplied
<COORDINATE_FUNCTION> +
<FACET_FUNCTION> +
<SCALE_FUNCTION> +
<THEME_FUNCTION>
```

Example:
```
ggplot(data = mpg) +
   geom_point(mapping = aes(x = cty, y = hwy))
```


Example:
```
ggplot(data = mpg) +
   geom_point(mapping = aes(x = cty, y = hwy))
```



Identical plot generated using shorthand of excluding argument names:
```
ggplot(mpg) + geom_point(aes(x = cty, y = hwy))
```

Class Walkthrough

```
## set your working directory to your class
## folder:
## Session -> Set Working Directory -> Choose Directory

classURL =
"github.com/flyaflya/buad621/archive/master.zip“

usethis::use_course(url = classURL, destdir = getwd())
```


Class Walkthrough: Four Steps of Data Visualization – Example – Coors Field

1) Purpose – Coors field is rumored to be the easiest field to score runs at. Is it true?
2) Content – what information should be shown?
3) Structure – how do we visualize the content?
4) Formatting – making it more informative, persuasive, and/or aesthetic?

Open `coorsField.R` (make sure `baseball.Rdata` is in your working directory)

Working in layers

(see `hflightsDelay.r` for R-code if interested)

- Let's use the hflights data again
  - PURPOSE: Assume we want to investigate delay by day of week and by carrier
  - Step 1: Get the dataframe of **data** that you need

```
delayByCarrier =
select(.data=hflights, DayOfWeek, UniqueCarrier, ArrDelay)
```