Agenda

- Exploratory Data Analysis
- Presentation
- Critique
- Coding
- Syllabus

What is data visualization?

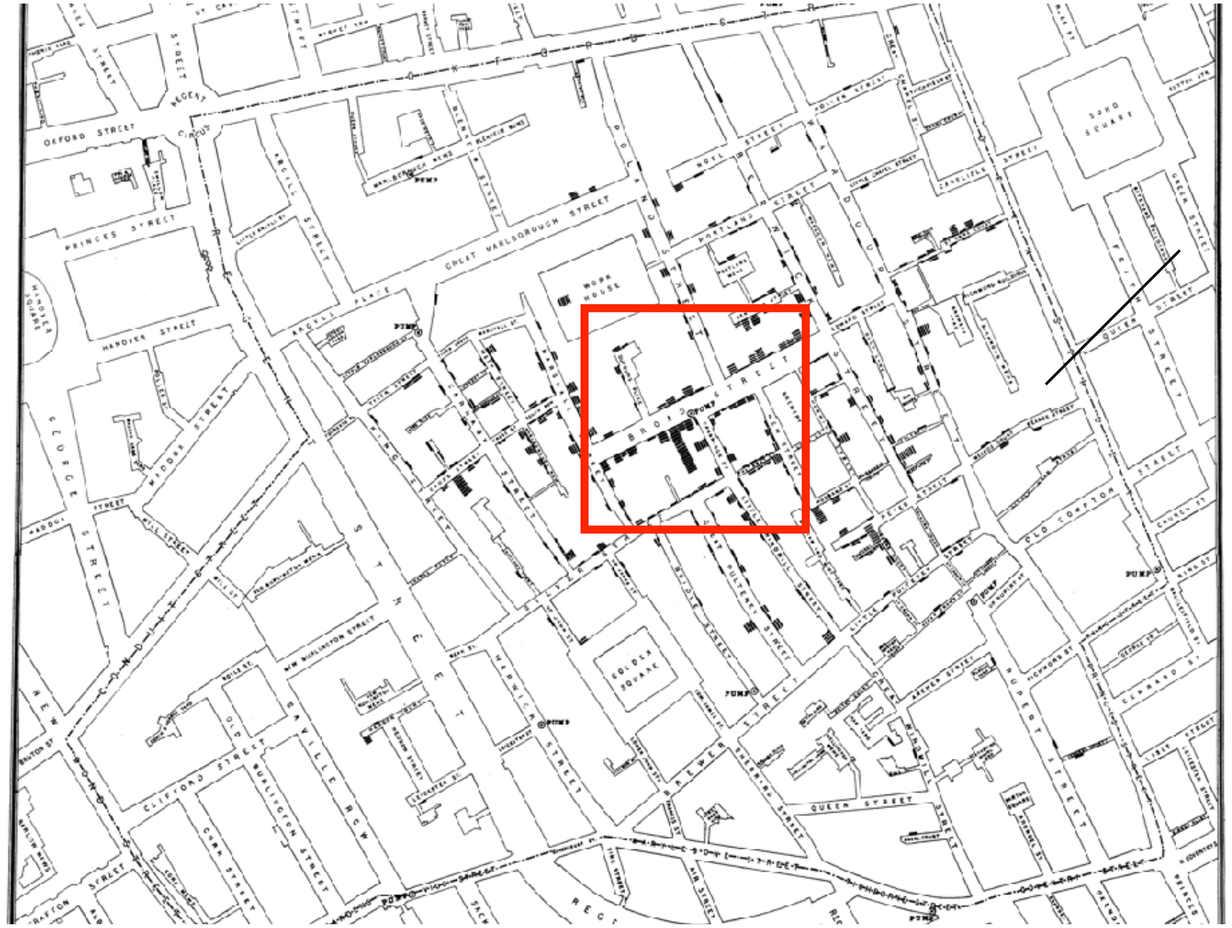
- relatively new field (but long history)
- multidisciplinary
- lack of consensus

EXPLORATORY DATA ANALYSIS

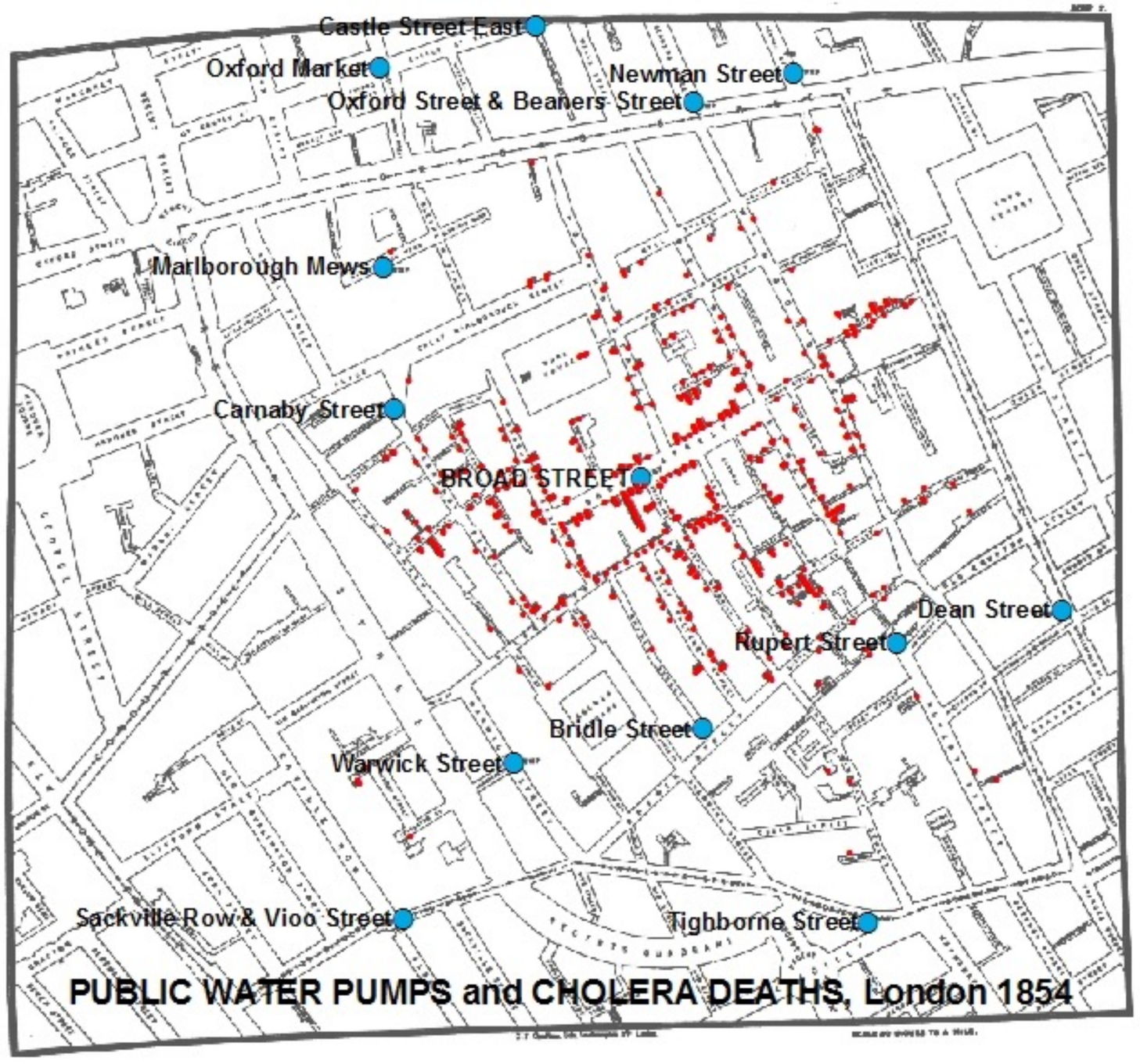
Our Goal: Data Insight

- Either for us, as data analysts, or for an audience
- Not all data visualizations produce insights
- Some are beautiful
- Some misinform (deliberately or not)
- Focus should be on the data, not the tool

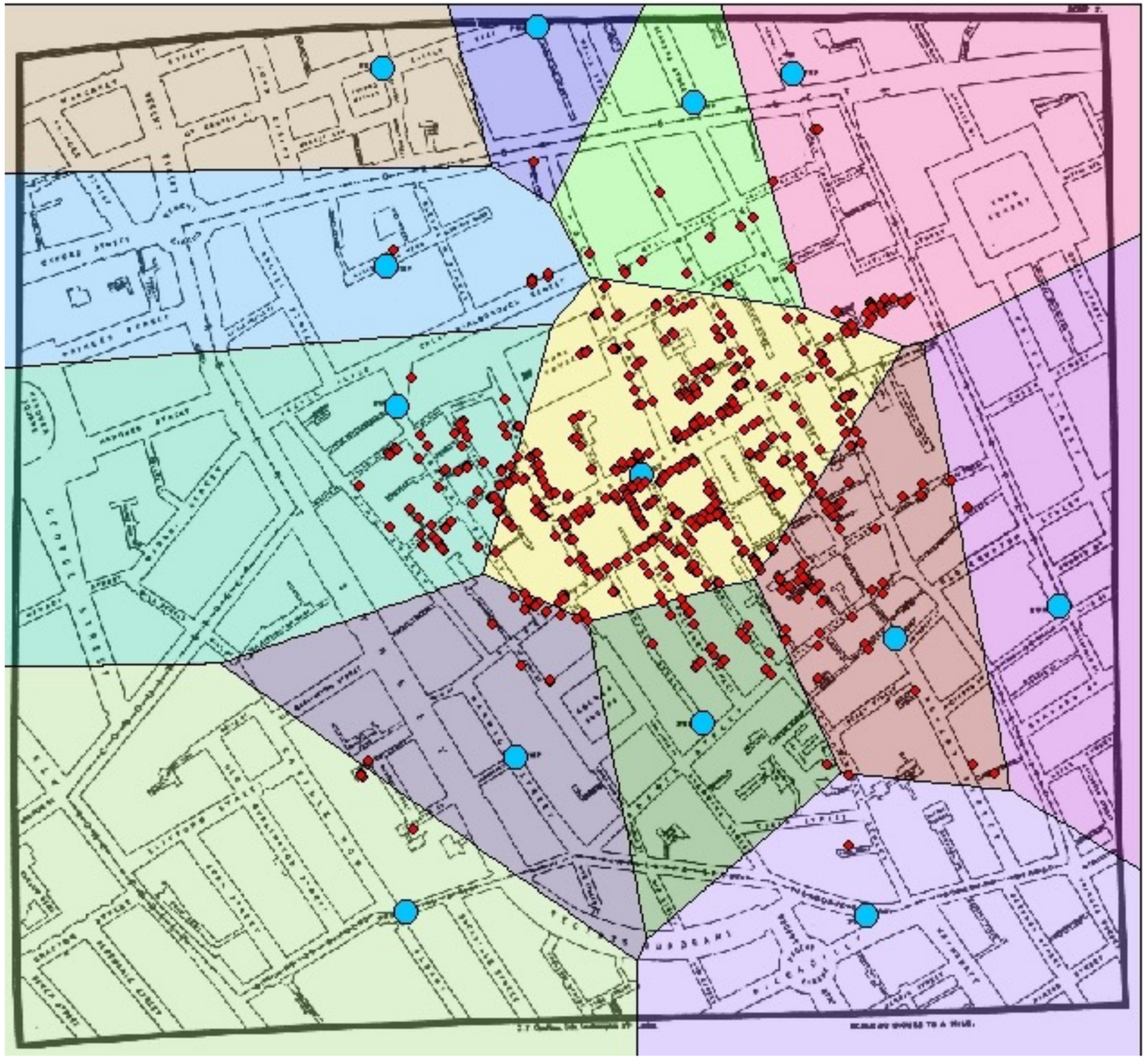
John Snow, Cholera Map 1854



Broad St. pump


Castle Street East
Oxford Market
Newman Street
Oxford Street & Beaners Street
Marlborough Mews
Carnaby Street
BROAD STREET
Dean Street
Rupert Street
Bridle Street
Warwick Street
Sackville Row & Vioo Street
Tighborne Street
PUBLIC WATER PUMPS and CHOLERA DEATHS, London 1854

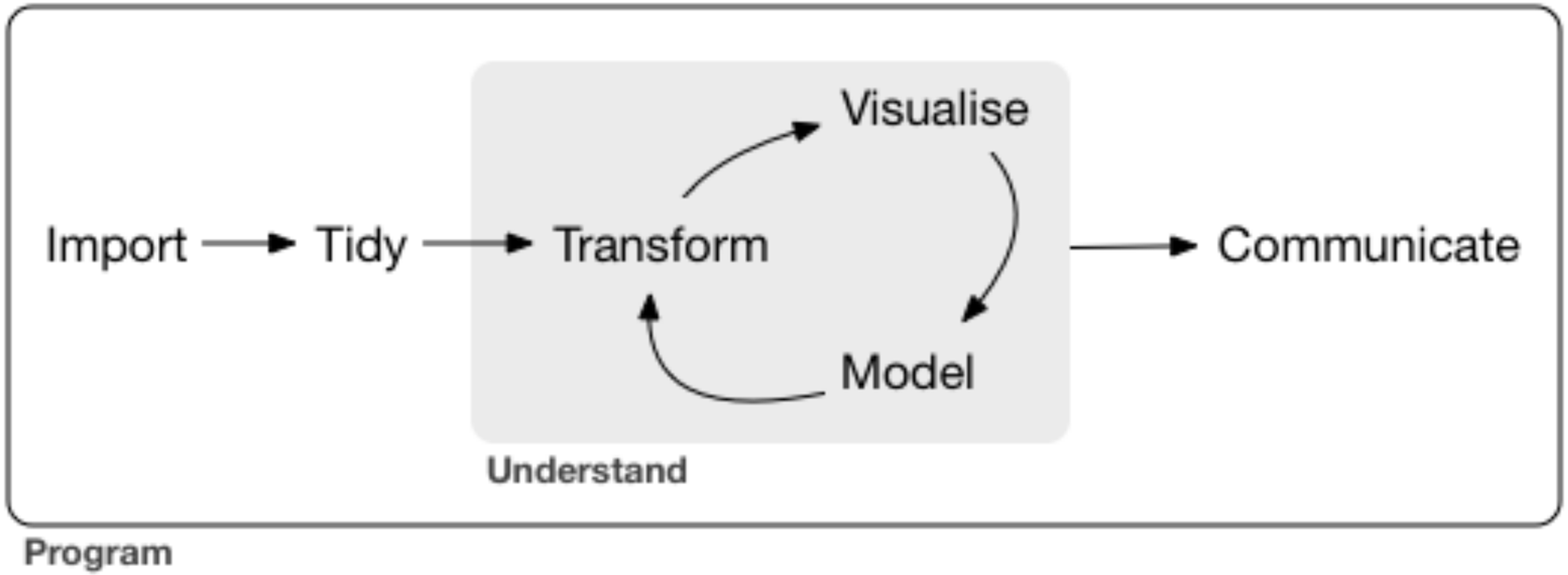
Thiessen polygons



Exploratory Data Analysis

- Data visualization for data science
  - detecting patterns
  - finding outliers
  - making comparisons
  - identifying clusters

Data Science Process



"Visualization is a fundamentally human activity."

Anscombe's Quartet

x1	x2	x3	x4	y1	y2	y3	y4
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	8	7.24	6.13	6.08	5.25
4	4	4	19	4.26	3.10	5.39	12.50
12	12	12	8	10.84	9.13	8.15	5.56
7	7	7	8	4.82	7.26	6.42	7.91
5	5	5	8	5.68	4.74	5.73	6.89

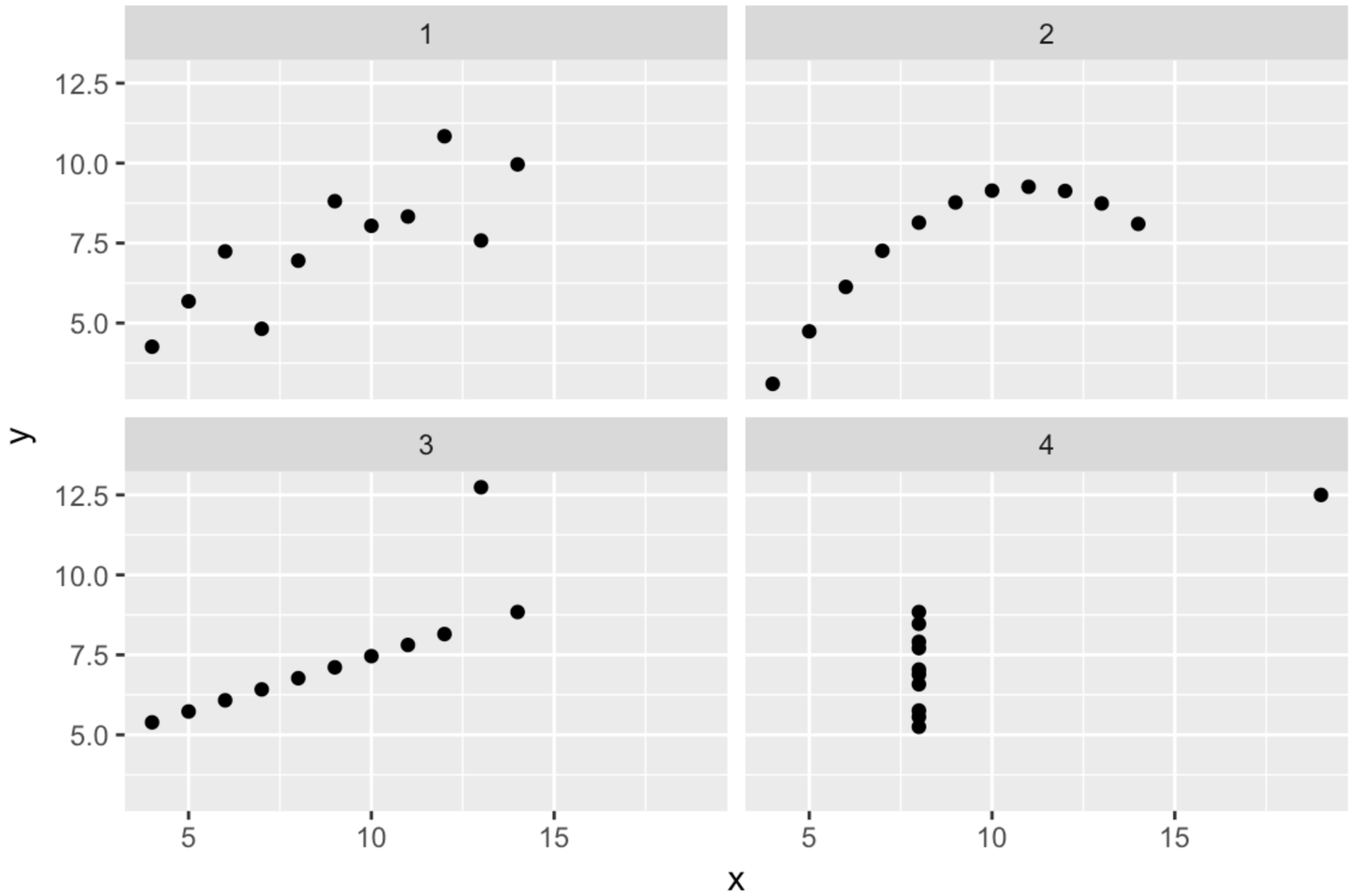
Numeric summary

Each of the four data sets yields the same standard output from a typical regression program, namely

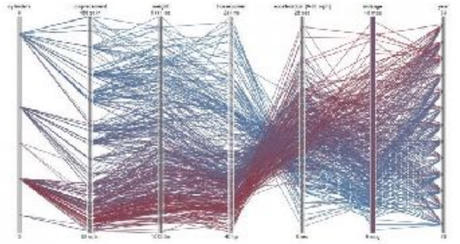
Number of observations $(n) = 11$
Mean of the x's $(\bar{x}) = 9.0$
Mean of the y's $(\bar{y}) = 7.5$
Regression coefficient (b_1) of y on $x = 0.5$
Equation of regression line: $y = 3 + 0.5\,x$
Sum of squares of $x - \bar{x} = 110.0$
Regression sum of squares $= 27.50$ (1 d.f.)
Residual sum of squares of $y = 13.75$ (9 d.f.)
Estimated standard error of $b_1 = 0.118$
Multiple $R^2 = 0.667$

F.J. Anscombe, "Graphs in Statistical Analysis"

Graphical summary



How do we gain insight?



- Deep understanding of the dataset, where it came from, what its limitations are
- Experiment with different graphic forms, based on theory on what forms work well with different data types

Interactivity

- For the data analyst, visualizing data is always an interactive process
- Choice whether we allow the users to explore the dataset or present them with the "best shots"
- Graphs are easy to create, anyone can do it, but not always well

Titantic Dataset (R)

```r
str(Titanic)
```

```
##  table [1:4, 1:2, 1:2, 1:2] 0 0 35 0 0 0 17 0 118 154 .
##  - attr(*, "dimnames")=List of 4
##   ..$ Class   : chr [1:4] "1st" "2nd" "3rd" "Crew"
##   ..$ Sex     : chr [1:2] "Male" "Female"
##   ..$ Age     : chr [1:2] "Child" "Adult"
##   ..$ Survived: chr [1:2] "No" "Yes"
```

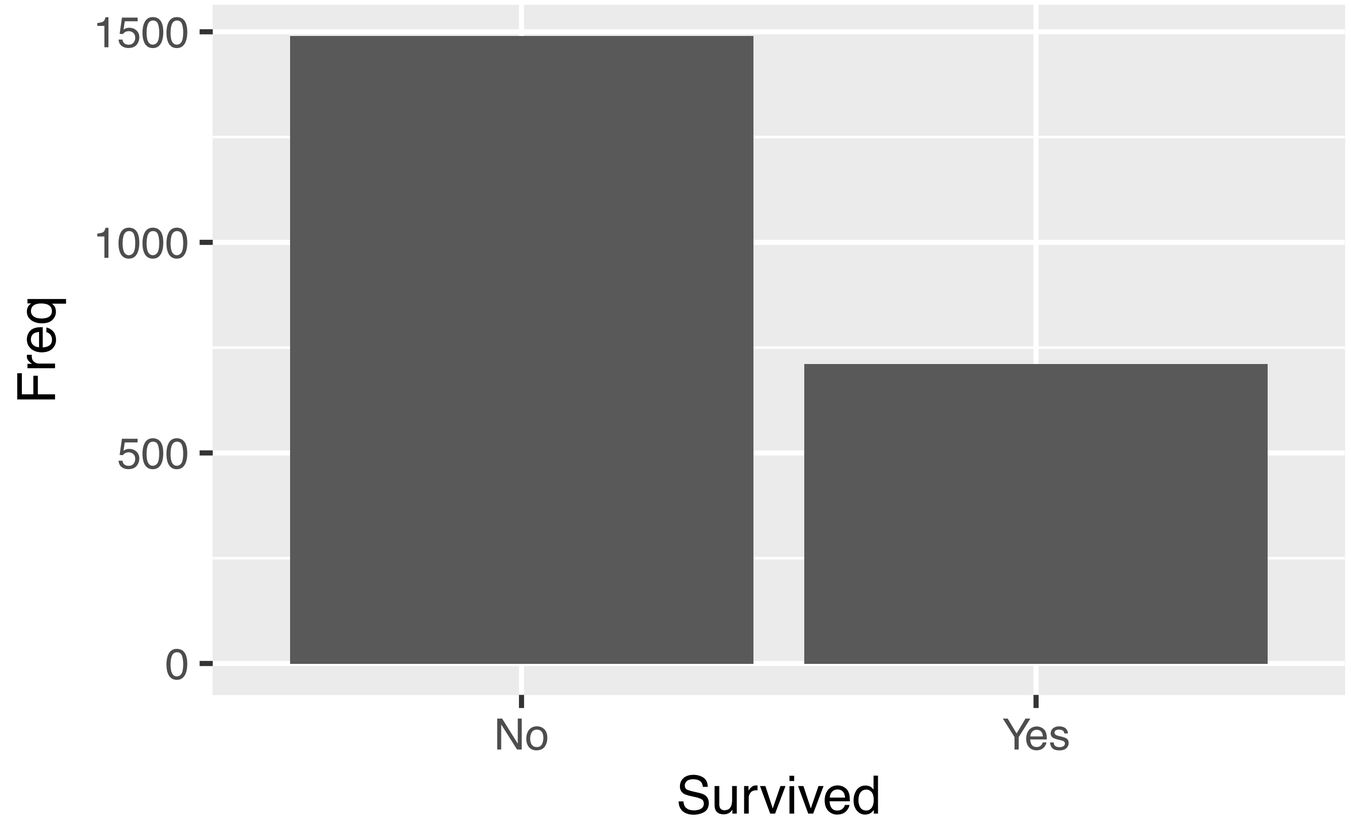
```r
df <- data.frame(Titanic)
str(df)
```

```
## 'data.frame':    32 obs. of  5 variables:
##  $ Class   : Factor w/ 4 levels "1st","2nd","3rd",..: 1
##  $ Sex     : Factor w/ 2 levels "Male","Female": 1 1 1
##  $ Age     : Factor w/ 2 levels "Child","Adult": 1 1 1
##  $ Survived: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1
##  $ Freq    : num  0 0 35 0 0 0 17 0 118 154 ...
```

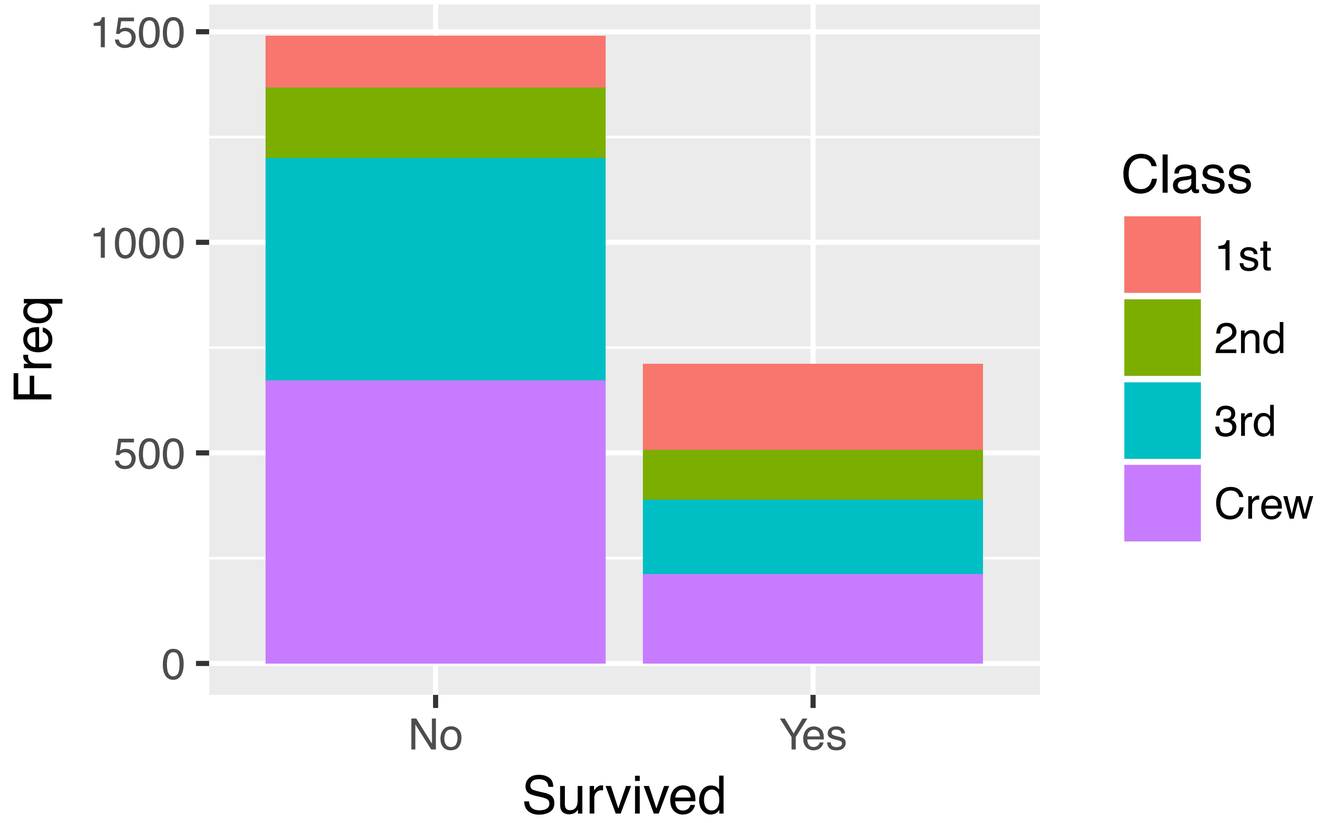
```r
head(df)
```

```
##   Class    Sex   Age Survived Freq
## 1   1st   Male Child       No    0
## 2   2nd   Male Child       No    0
## 3   3rd   Male Child       No   35
## 4  Crew   Male Child       No    0
## 5   1st Female Child       No    0
## 6   2nd Female Child       No    0
```

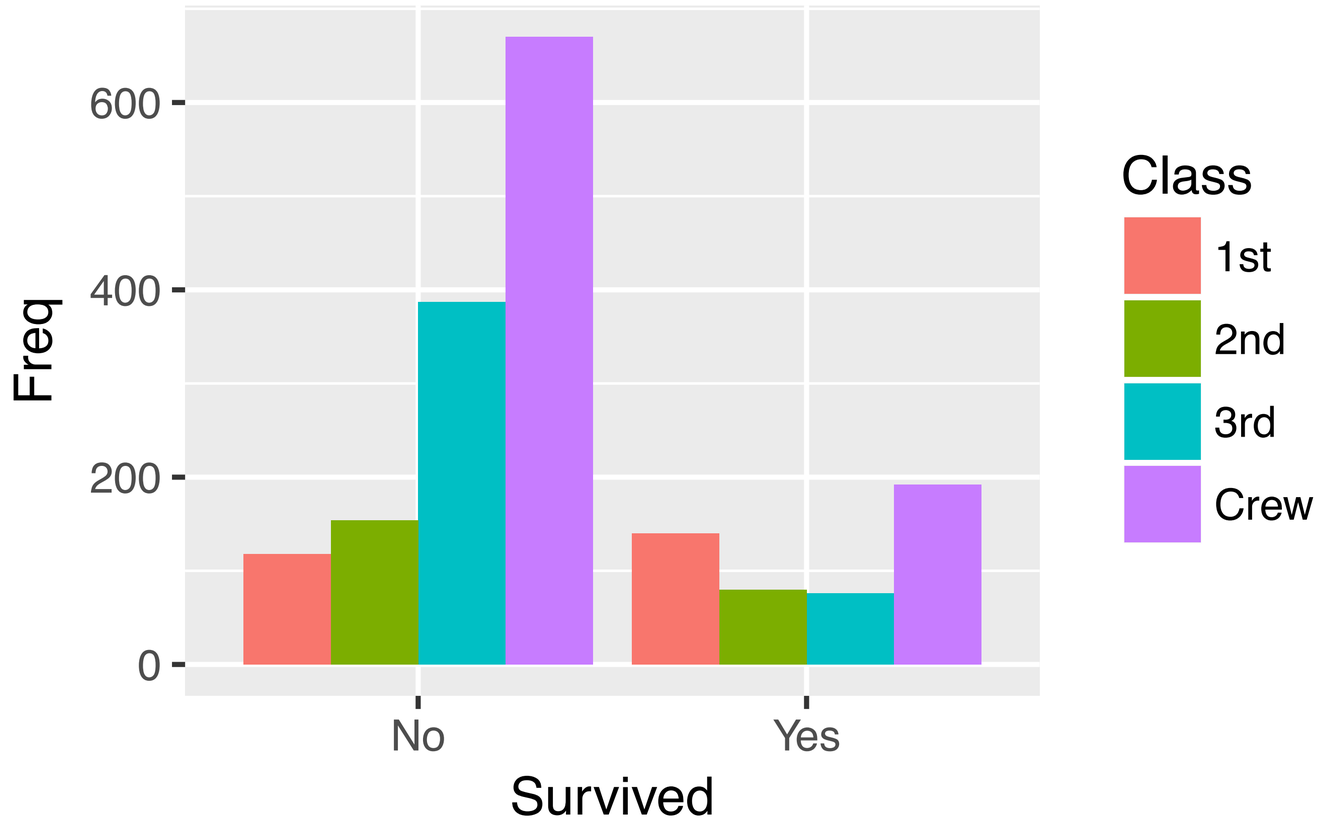
```r
library(ggplot2)
df <- data.frame(Titanic)
ggplot(df, aes(x = Survived, y = Freq)) + geom_bar(stat =
```



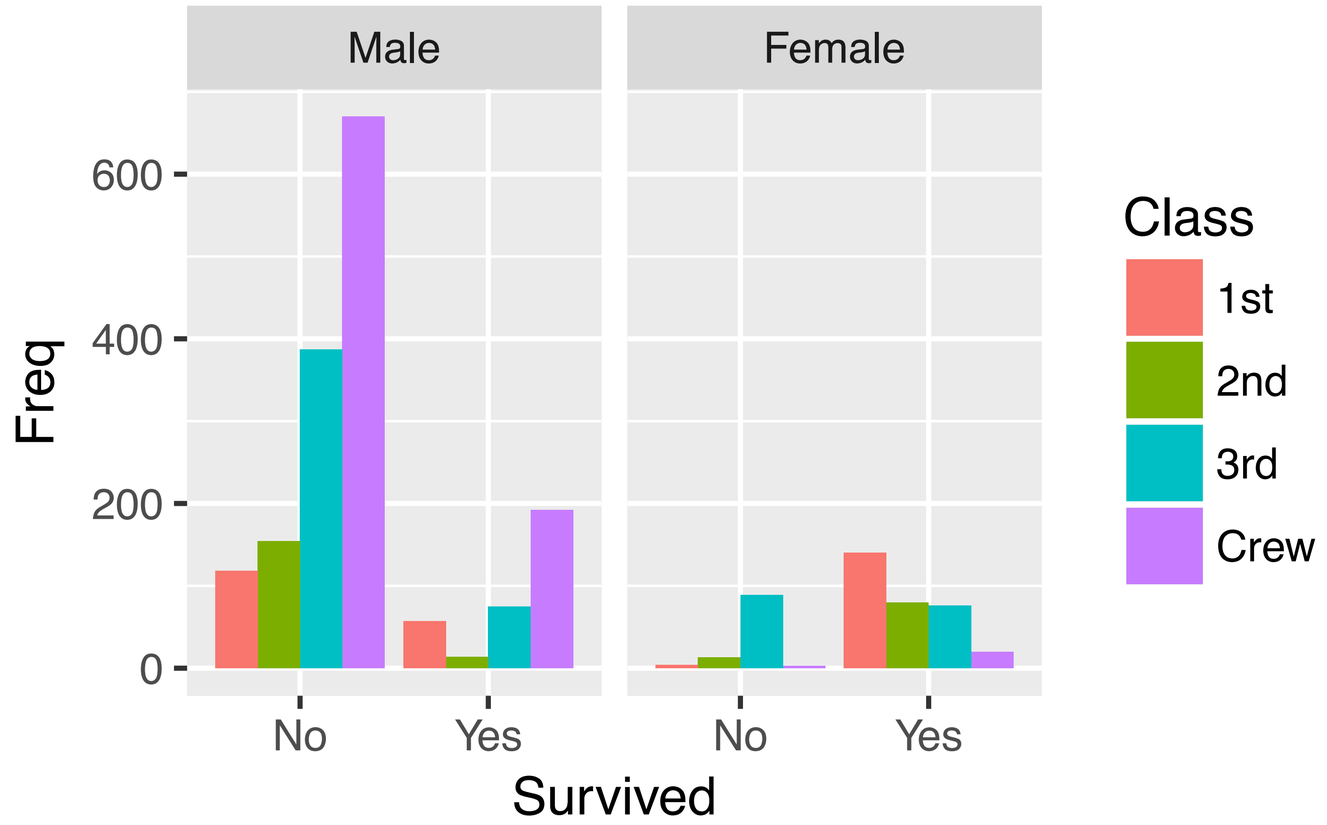
```
library(ggplot2)
df <- data.frame(Titanic)
ggplot(df, aes(x = Survived, y = Freq, fill = Class)) +
    geom_bar(stat = "identity")
```



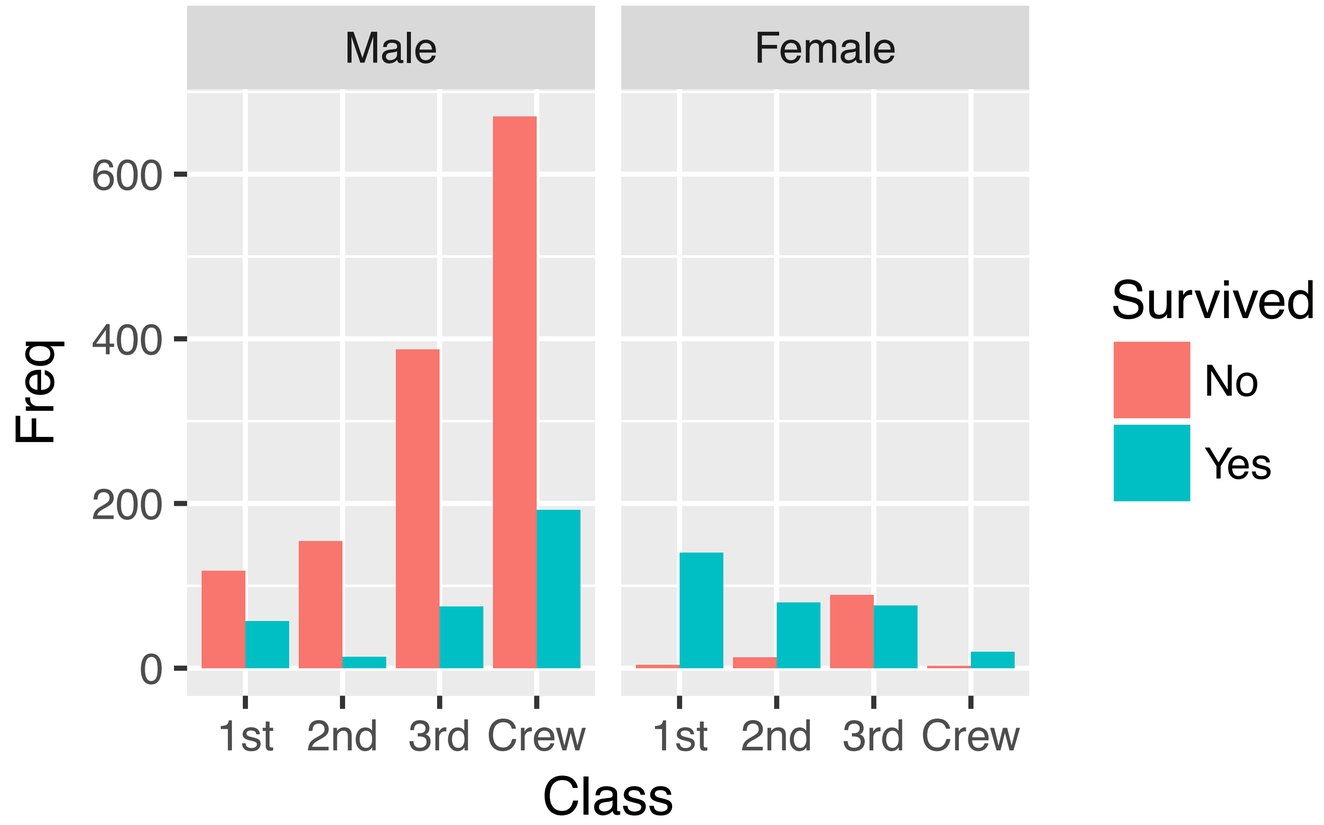
```r
library(ggplot2)
df <- data.frame(Titanic)
ggplot(df, aes(x = Survived, y = Freq, fill = Class)) +
    geom_bar(stat = "identity", position = "dodge")
```



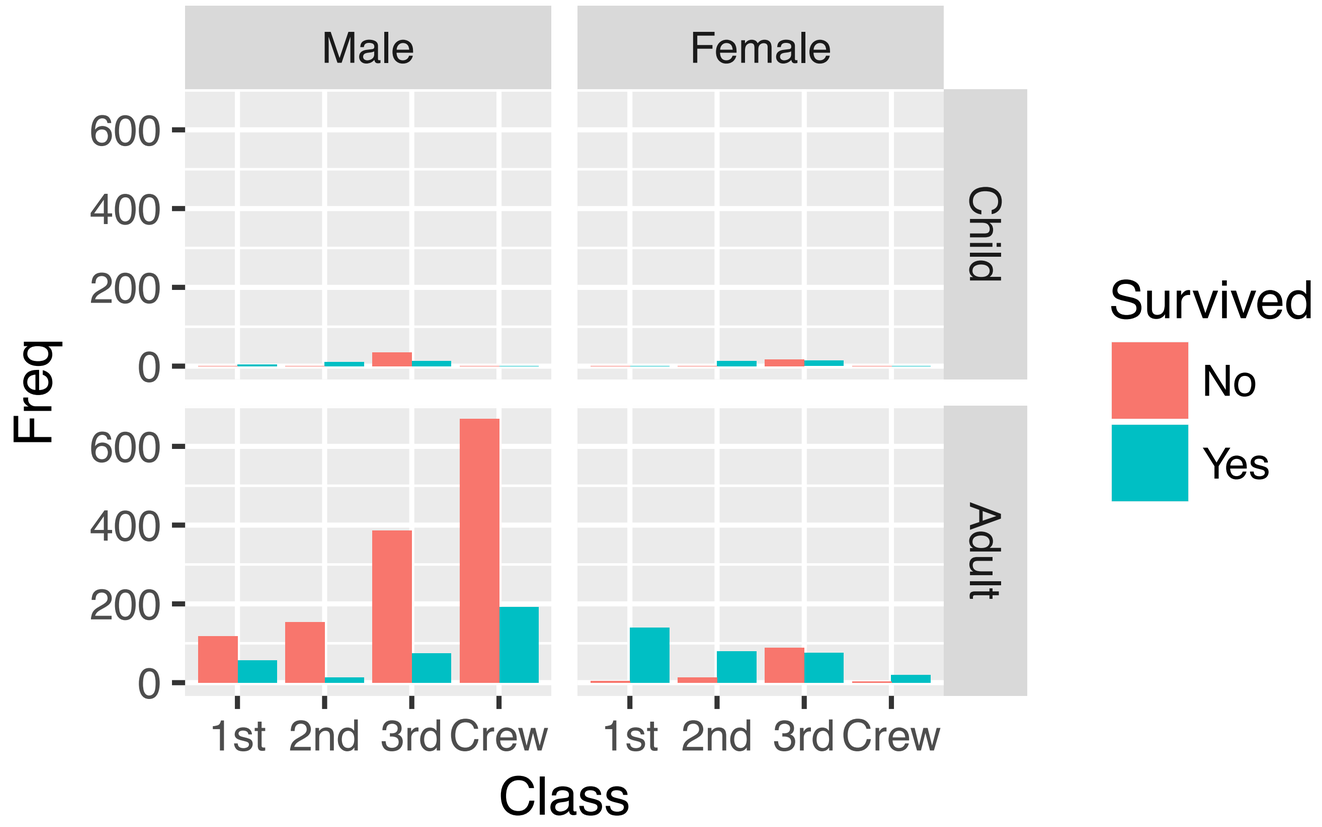
```r
library(ggplot2)
df <- data.frame(Titanic)
ggplot(df, aes(x = Survived, y = Freq, fill = Class)) +
    geom_bar(stat = "identity", position = "dodge") +
    facet_wrap(~Sex)
```

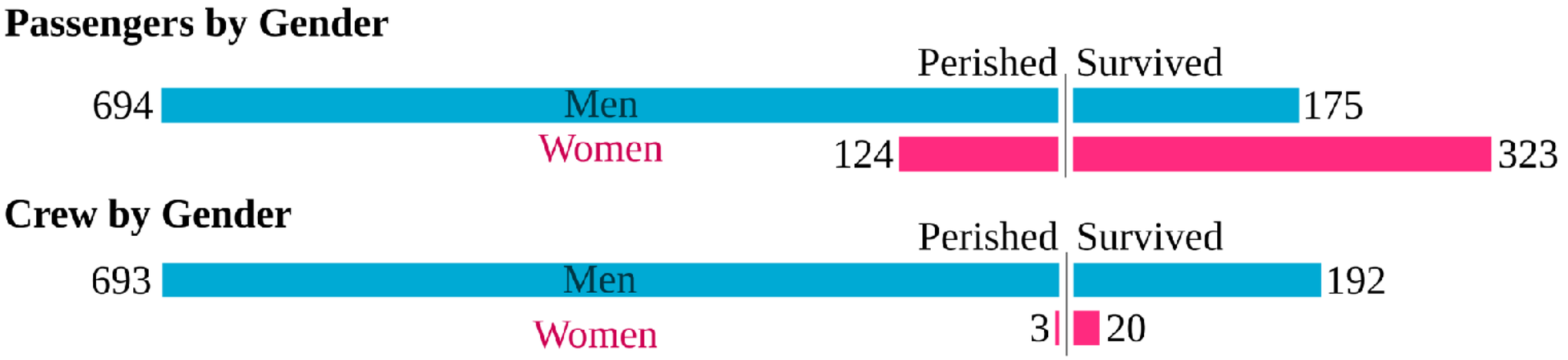


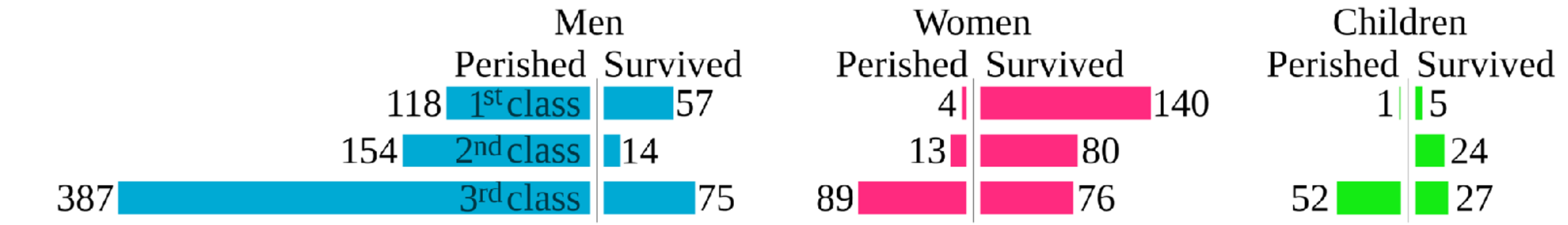
```r
library(ggplot2)
df <- data.frame(Titanic)
ggplot(df, aes(x = Class, y = Freq, fill = Survived)) +
    geom_bar(stat = "identity", position = "dodge") +
    facet_wrap(~Sex)
```



```r
library(ggplot2)
df <- data.frame(Titanic)
ggplot(df, aes(x = Class, y = Freq, fill = Survived)) +
    geom_bar(stat = "identity", position = "dodge") +
    facet_grid(Age~Sex)
```

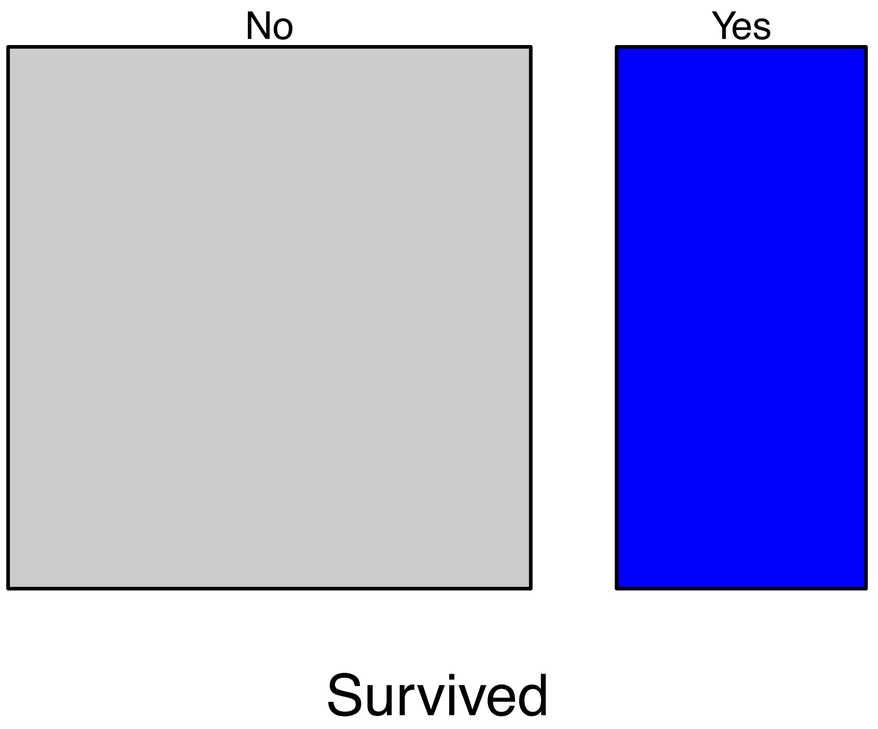




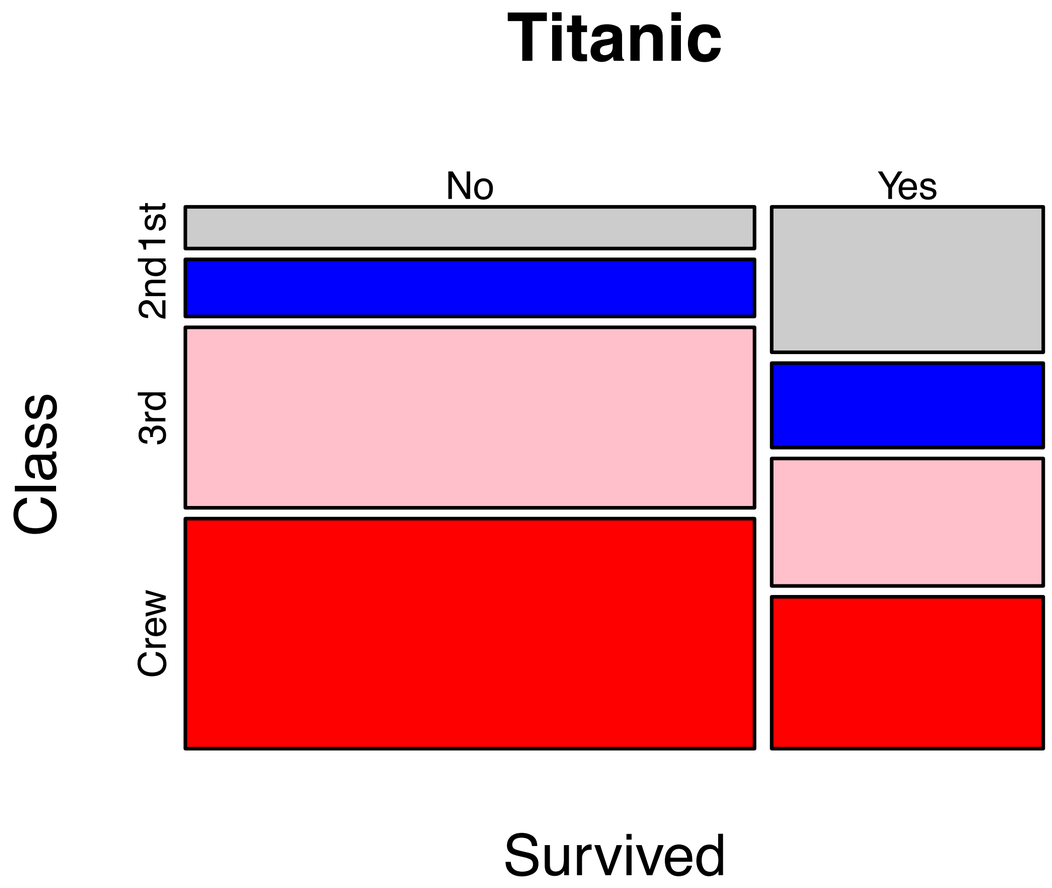


https://commons.wikimedia.org/wiki/File:Titanic_Survival_Stats.svg

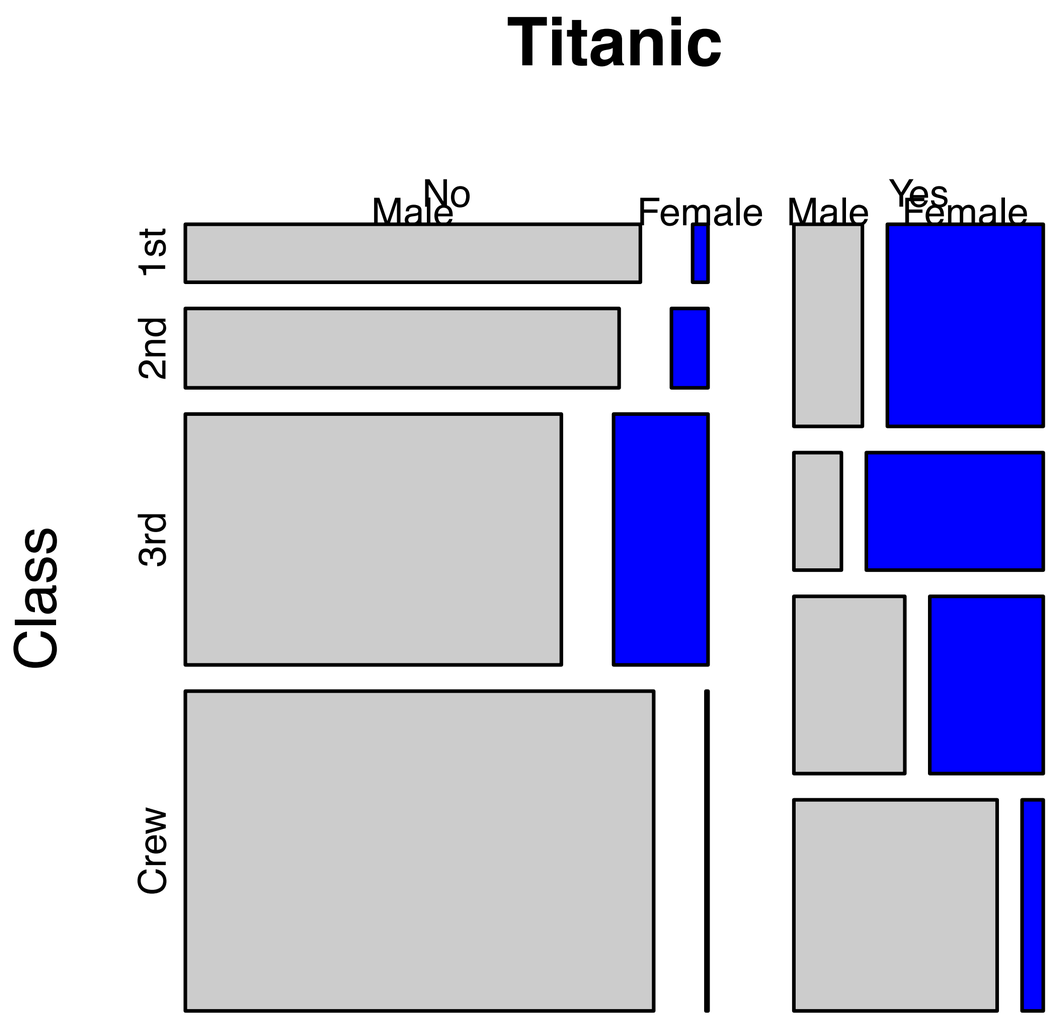
```
mosaicplot(~Survived, Titanic,
           color = c("grey80", "blue", "pink", "red"))
```



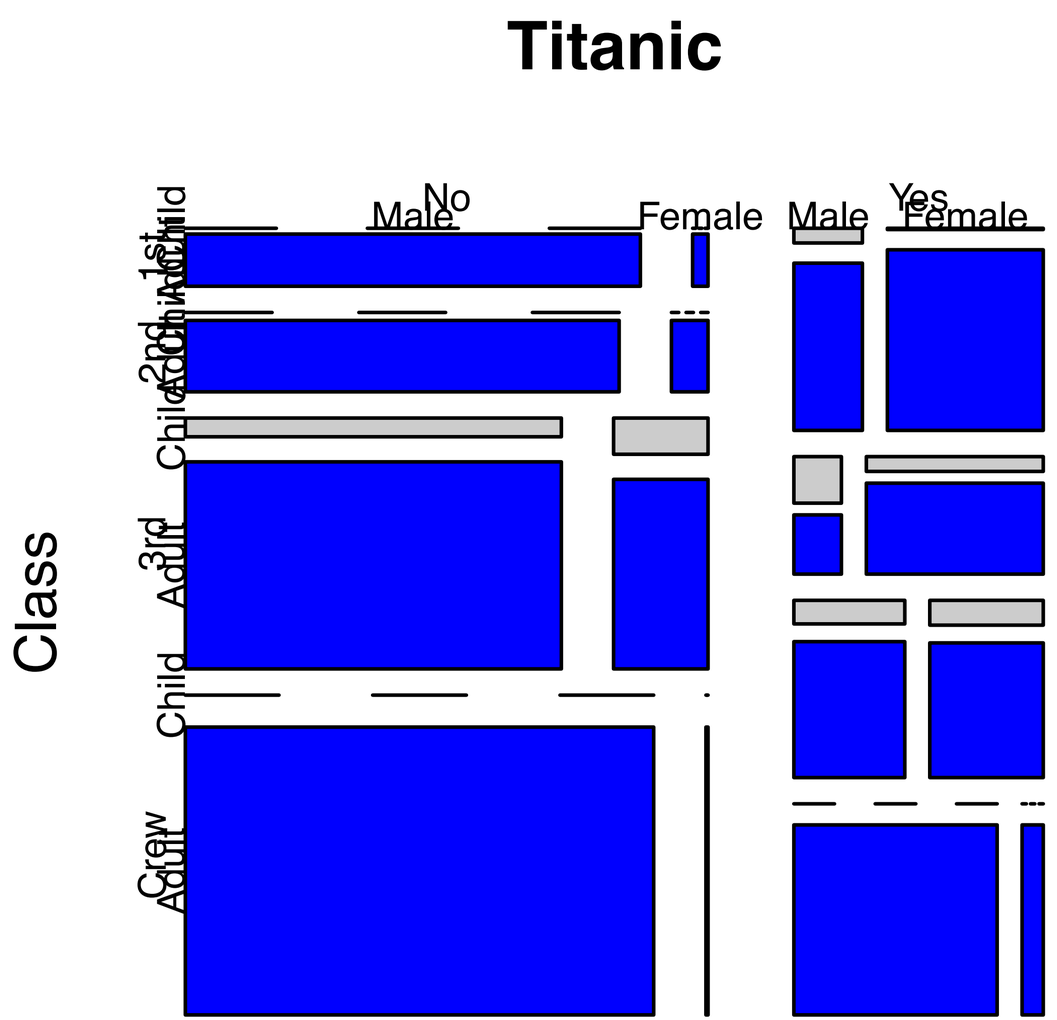
```
mosaicplot(~Survived + Class, Titanic,
           color = c("grey80", "blue", "pink", "red"))
```



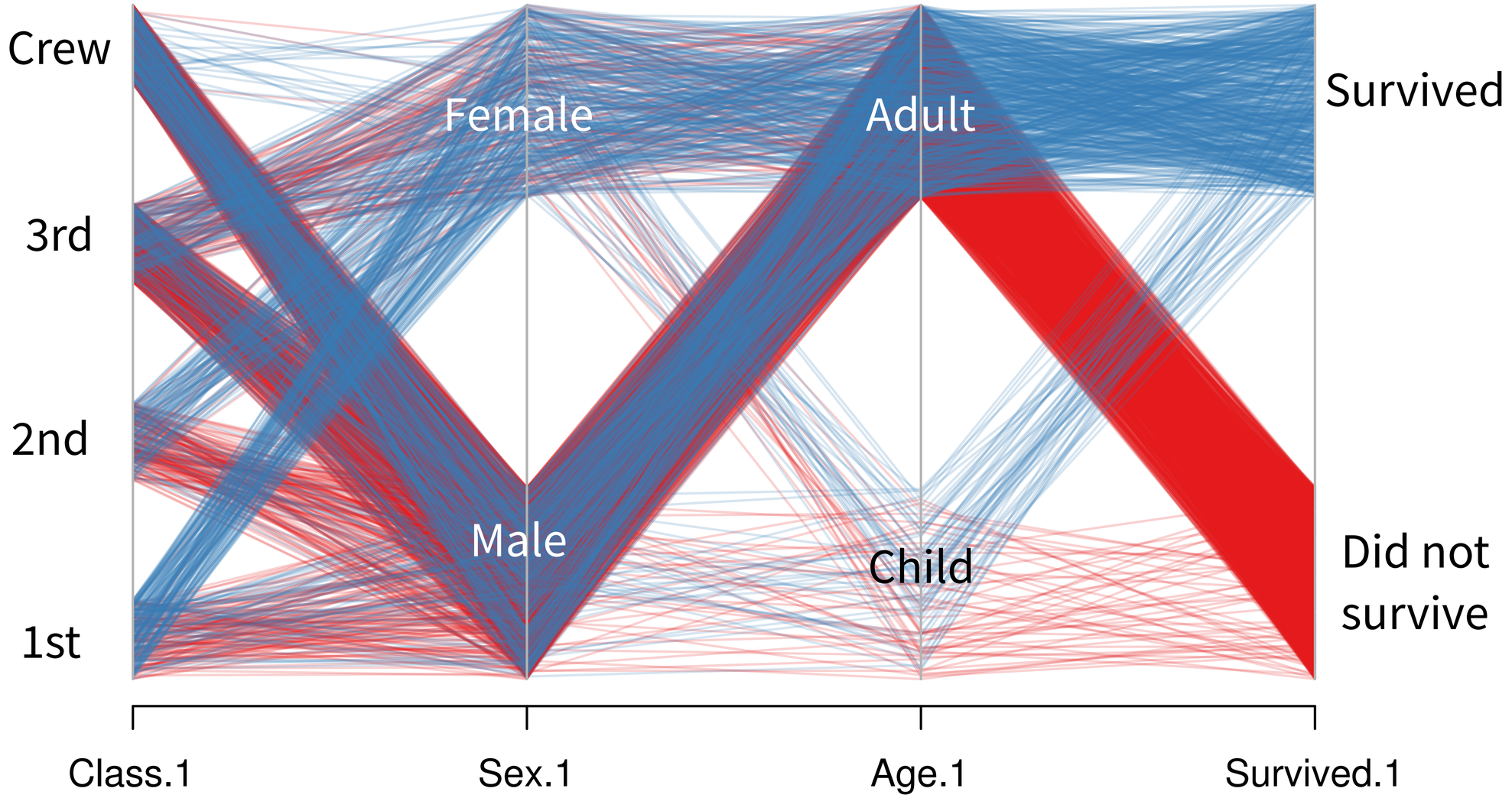
```
mosaicplot(~Survived + Class + Sex, Titanic,
           color = c("grey80", "blue", "pink", "red"))
```



```
mosaicplot(~Survived + Class + Sex + Age, Titanic,
           color = c("grey80", "blue", "pink", "red"))
```

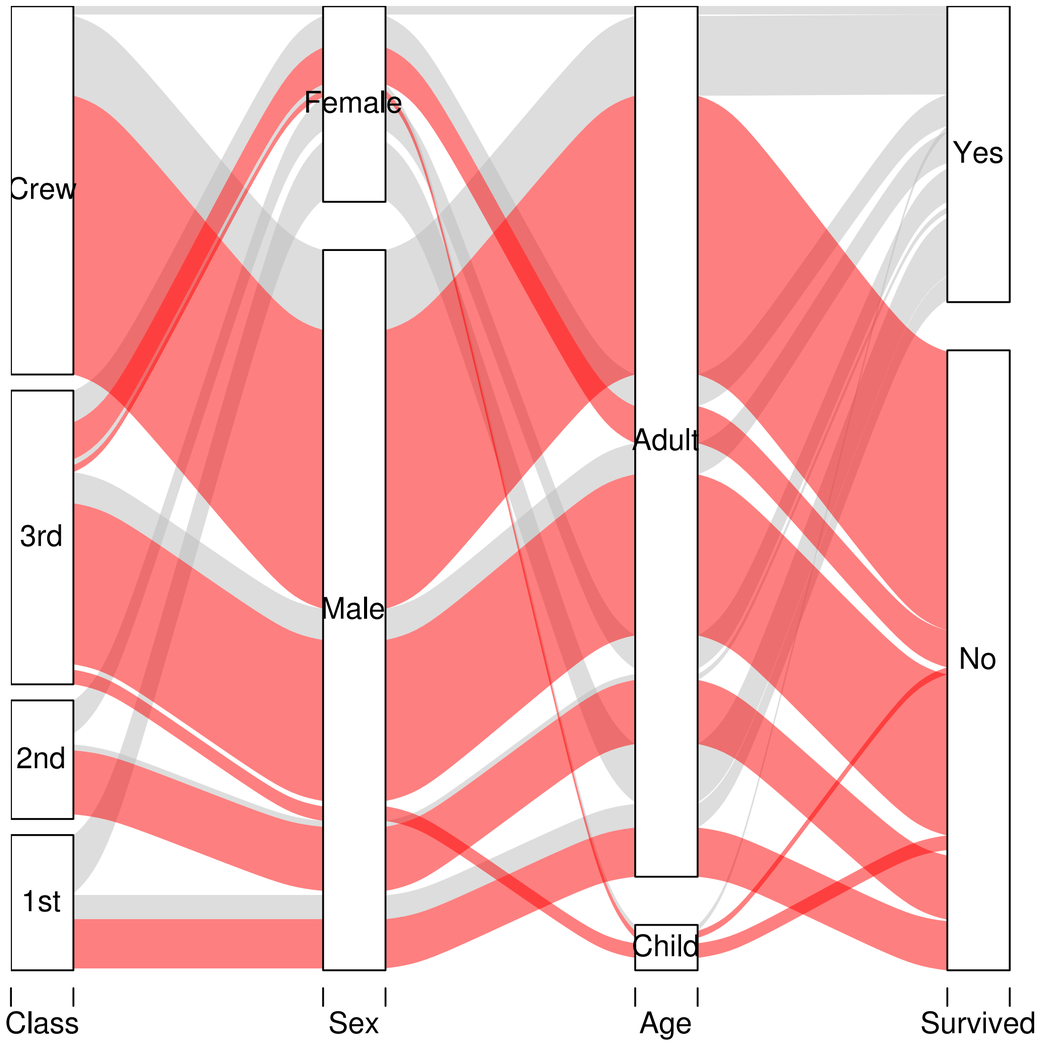


Parallel Coordinate Plot



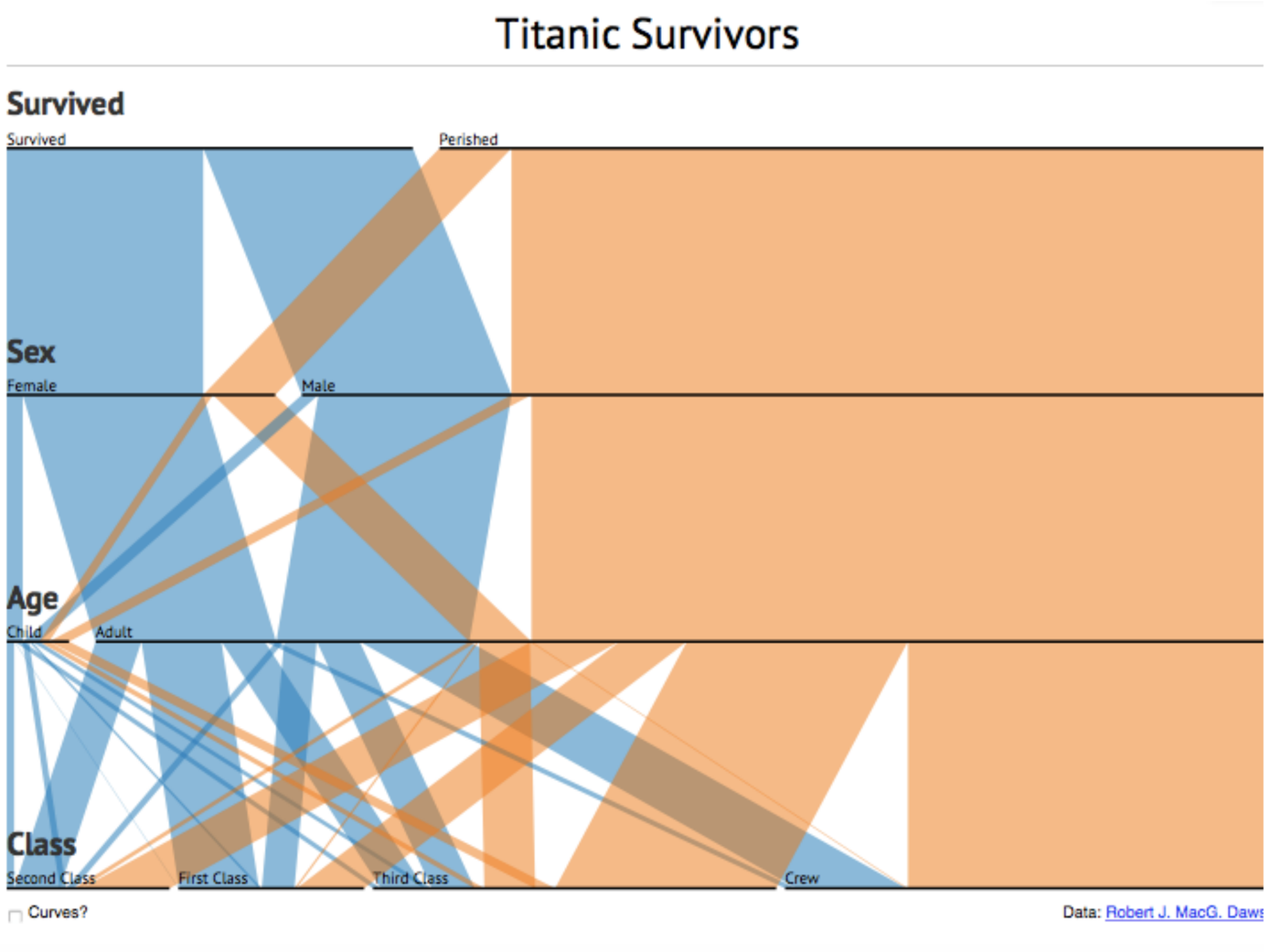
https://www.r-bloggers.com/alluvial-diagrams/

Alluvial Diagram



https://www.r-bloggers.com/alluvial-diagrams/

Interactive Parallel Sets



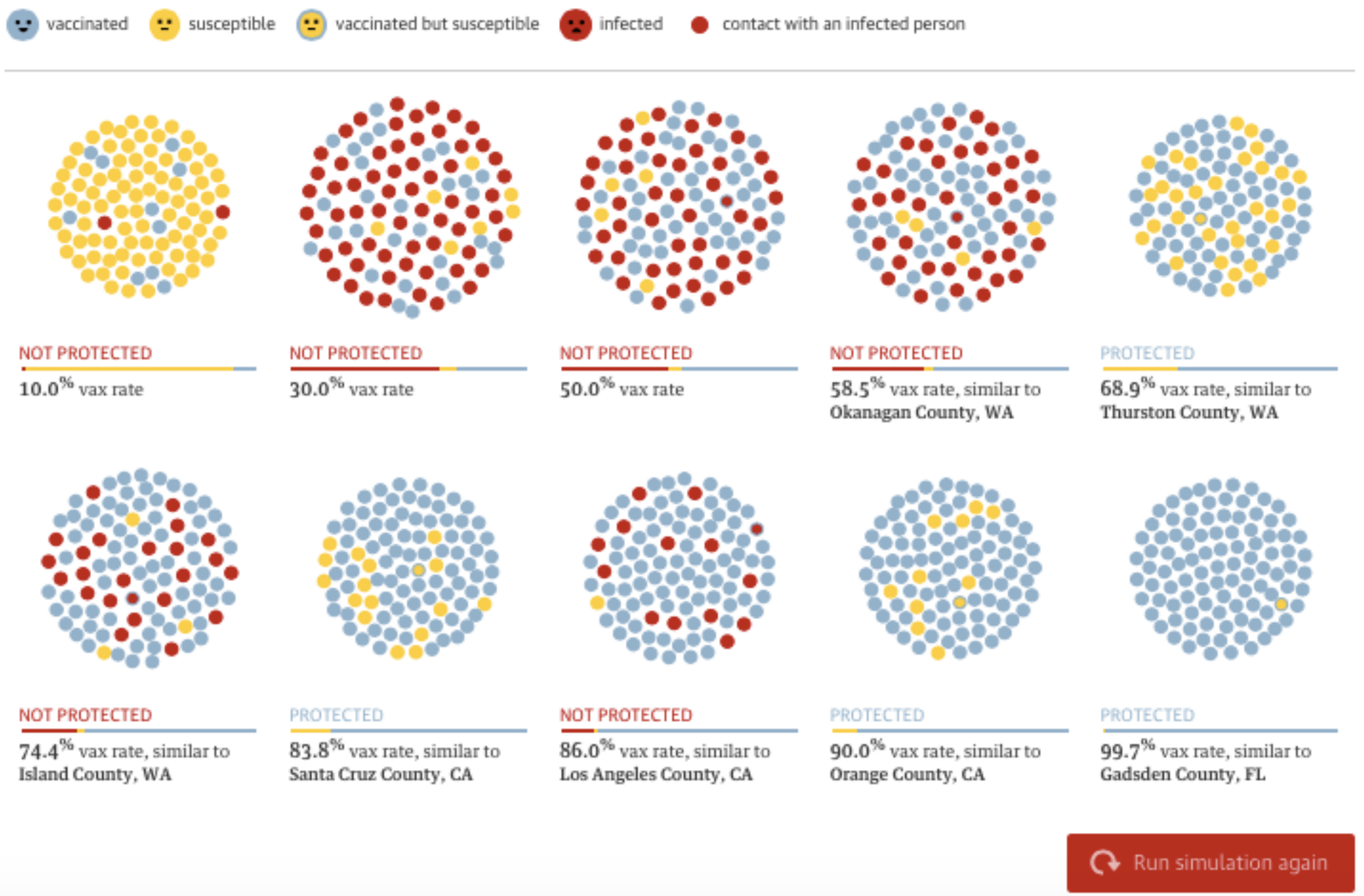
PRESENTATION

Exploration vs. Presentation

- Not mutually exclusive
- Visualizations that offer insight are likely to be shared
- Still, focus is different when exploratory a dataset for the first time vs. presenting to an audience, particularly a less technical one

55,272

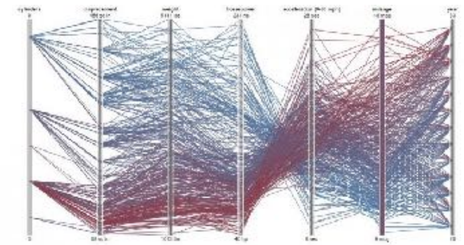
Watch how the measles outbreak spreads when kids get vaccinated – and when they don't



Interdisciplinary influences

Statistics

Subject Expertise

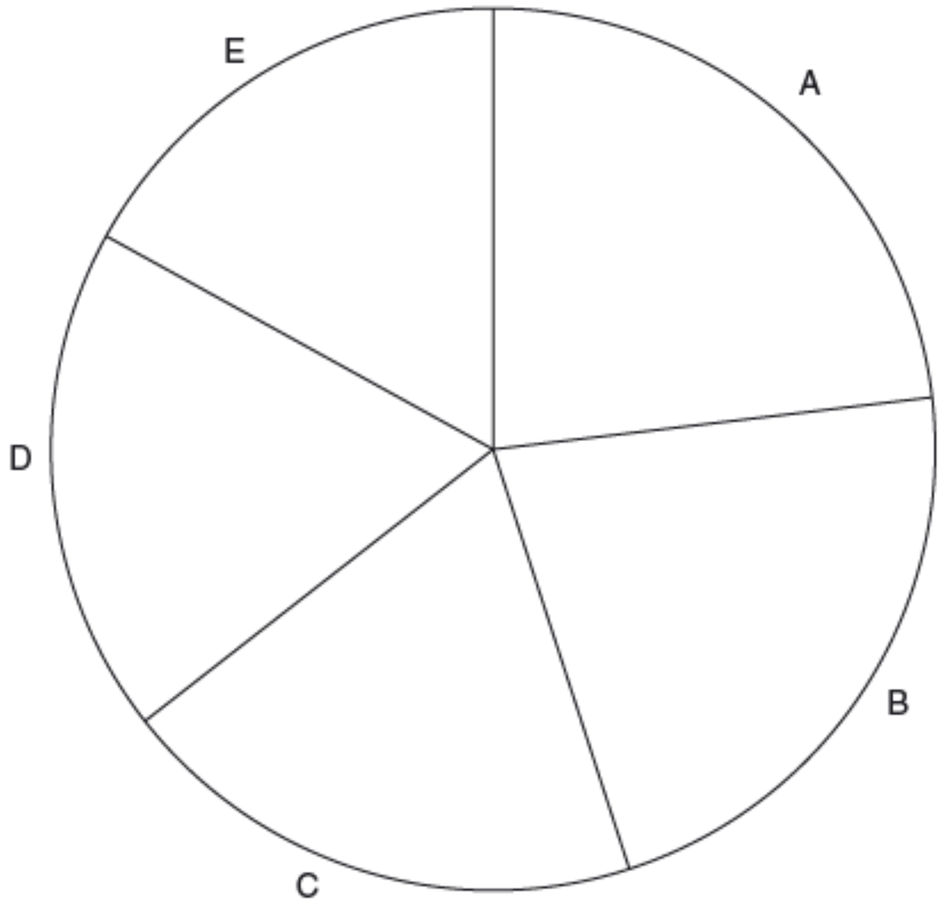


Graphic Design

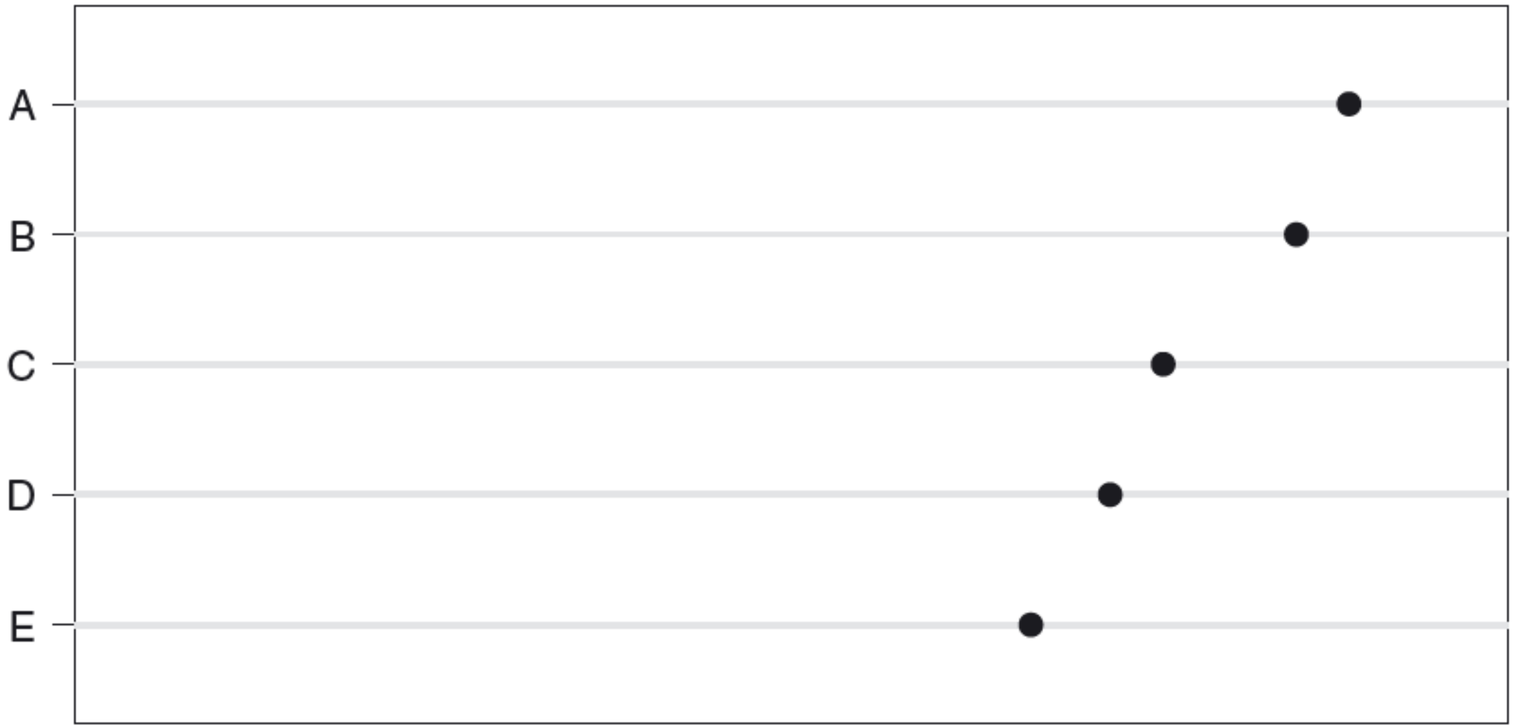
Psychology

Computer Science

Perception Studies



Cleveland Dot Plot



Graphics in R

pie {graphics} R Documentation

Pie Charts

Description

Draw a pie chart.

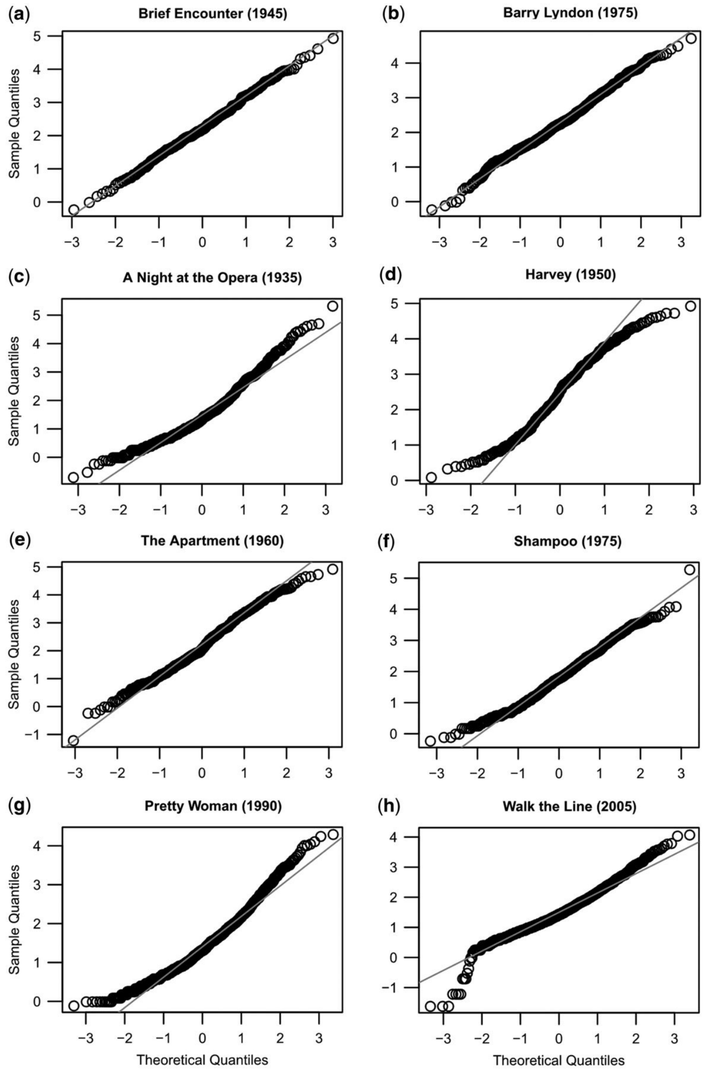
Usage

```
pie(x, labels = names(x), edges = 200, radius = 0.8,
    clockwise = FALSE, init.angle = if(clockwise) 90 else 0,
    density = NULL, angle = 45, col = NULL, border = NULL,
    lty = NULL, main = NULL, ...)
```

Arguments

`x`	a vector of non-negative numerical quantities. The values in `x` are displayed as the areas of pie slices.
`labels`	one or more expressions or character strings giving names for the slices. Other objects are coerced by `as.graphicsAnnot`. For empty or `NA` (after coercion to character) labels, no label nor pointing line is drawn.
`edges`	the circular outline of the pie is approximated by a polygon with this many edges.
`radius`	the pie is drawn centered in a square box whose sides range from *-1* to *1*. If the character strings labeling the slices are long it may be necessary to use a smaller radius.
`clockwise`	logical indicating if slices are drawn clockwise or counter clockwise (i.e., mathematically positive direction), the latter is default.

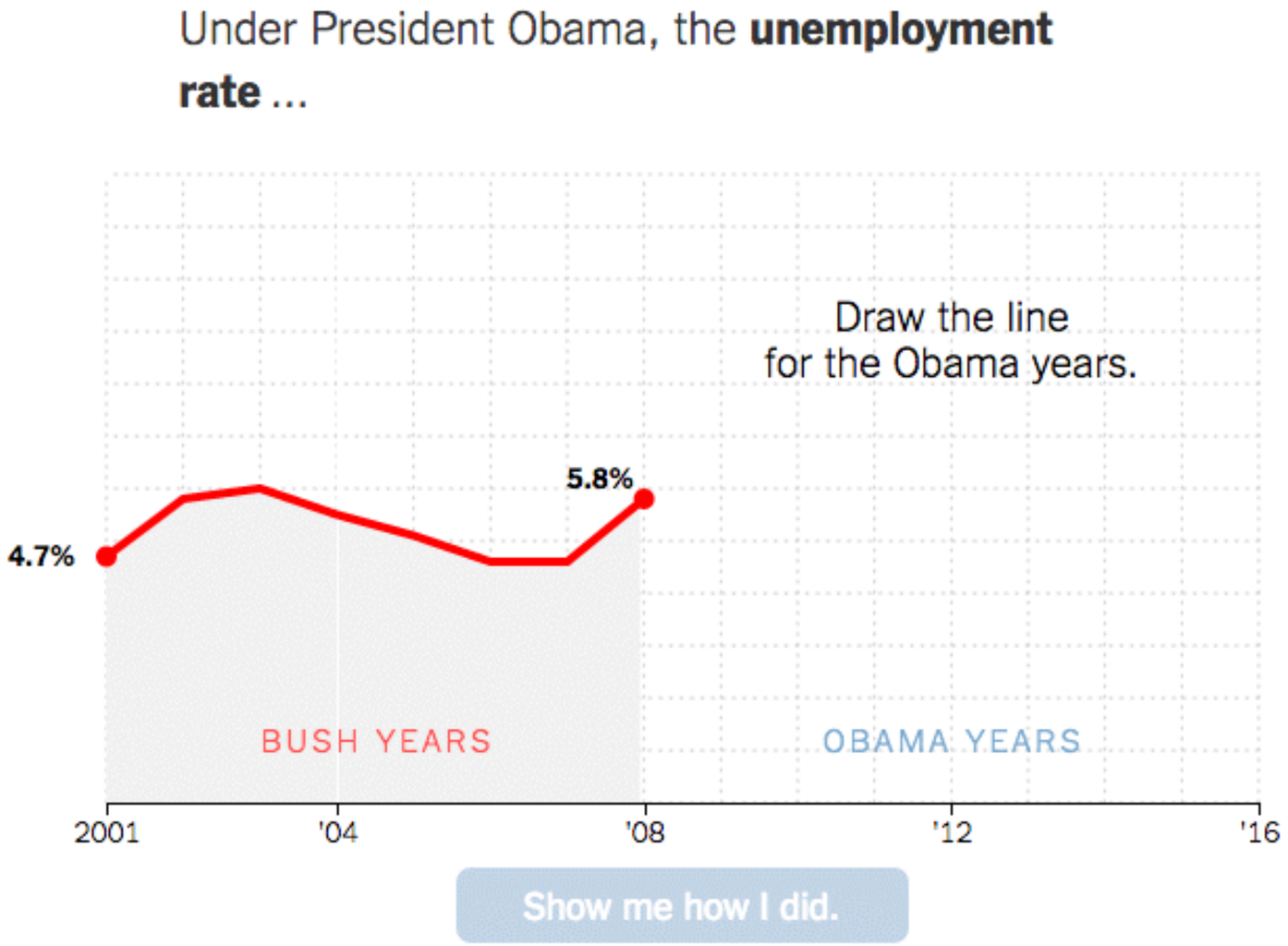
Audience



Normal probability plots of log-transformed shot lengths for eight films

http://dsh.oxfordjournals.org/content/30/1/137

You-draw-it


Under President Obama, the **unemployment rate** ...
Draw the line for the Obama years.
5.8%
4.7%
BUSH YEARS
OBAMA YEARS
2001
'04
'08
'12
'16
Show me how I did.

CRITIQUE

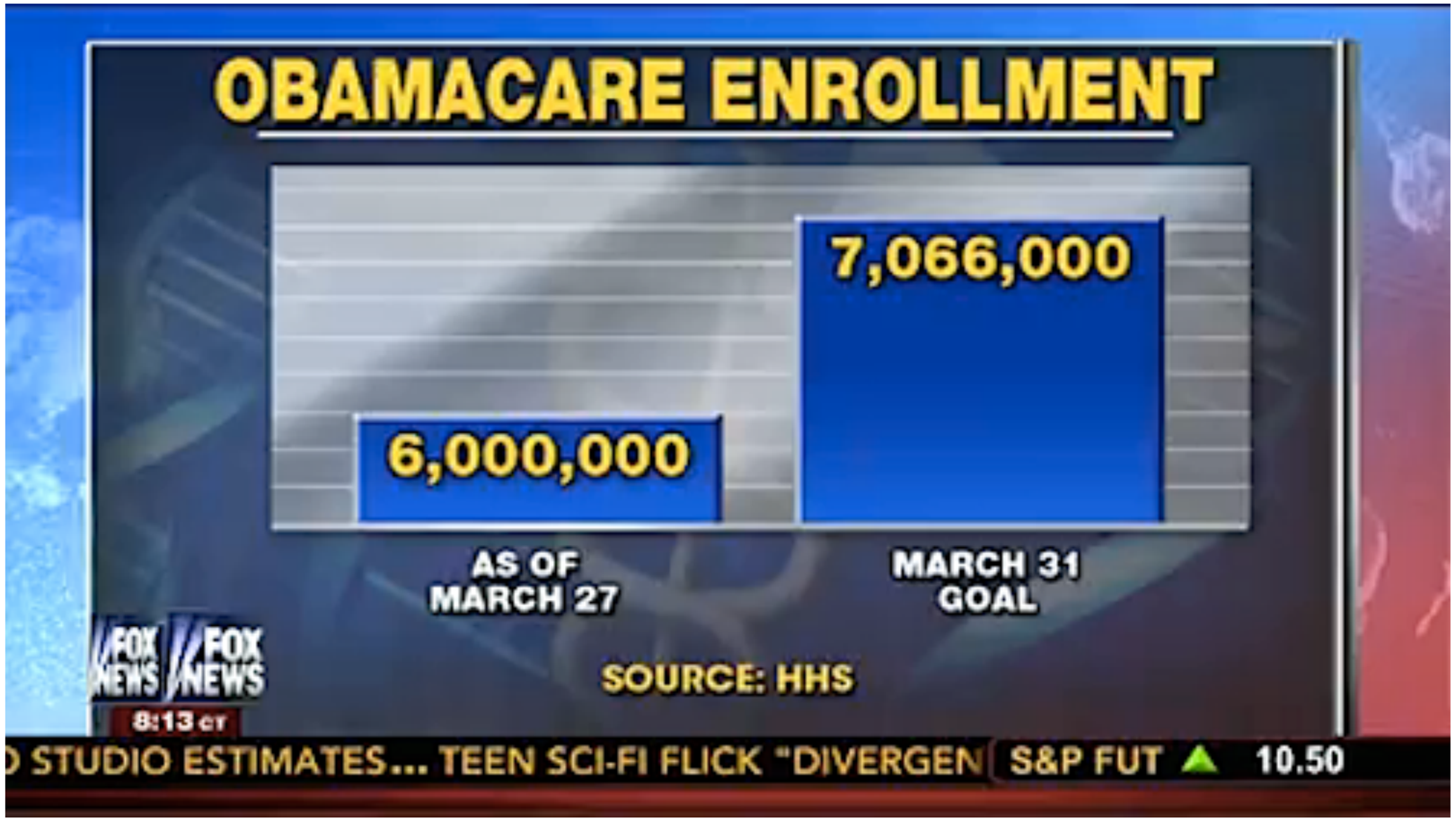
Growth of Data Visualization

"There is little evidence that the quality of the best graphics has improved over the last 100 years. I wonder if technology serves primarily as a **quantity**-multiplier, rather than a **quality**-multiplier." -Hadley Wickham

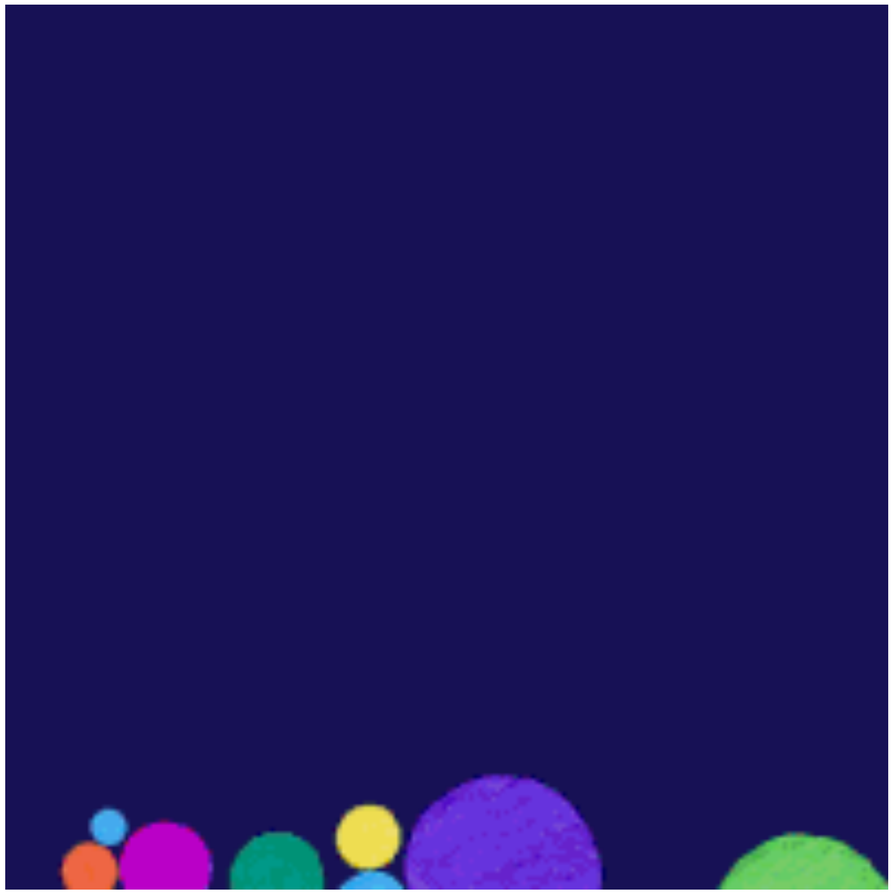
Evaluating Graphs

1. Wrong or misleading
2. Meaningless
3. Little added value
4. Good alternatives

Misleading Graph

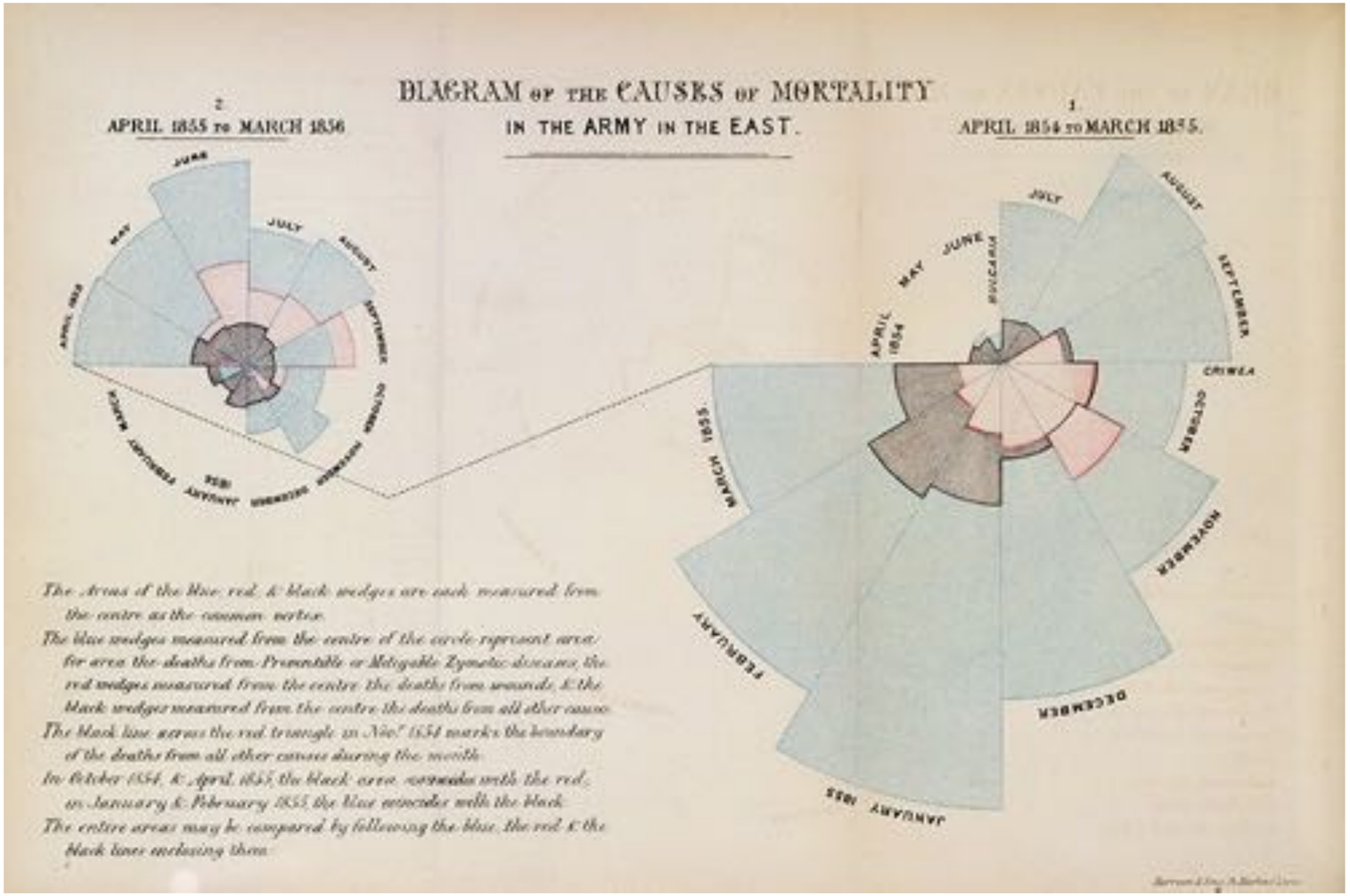


Data Visualization

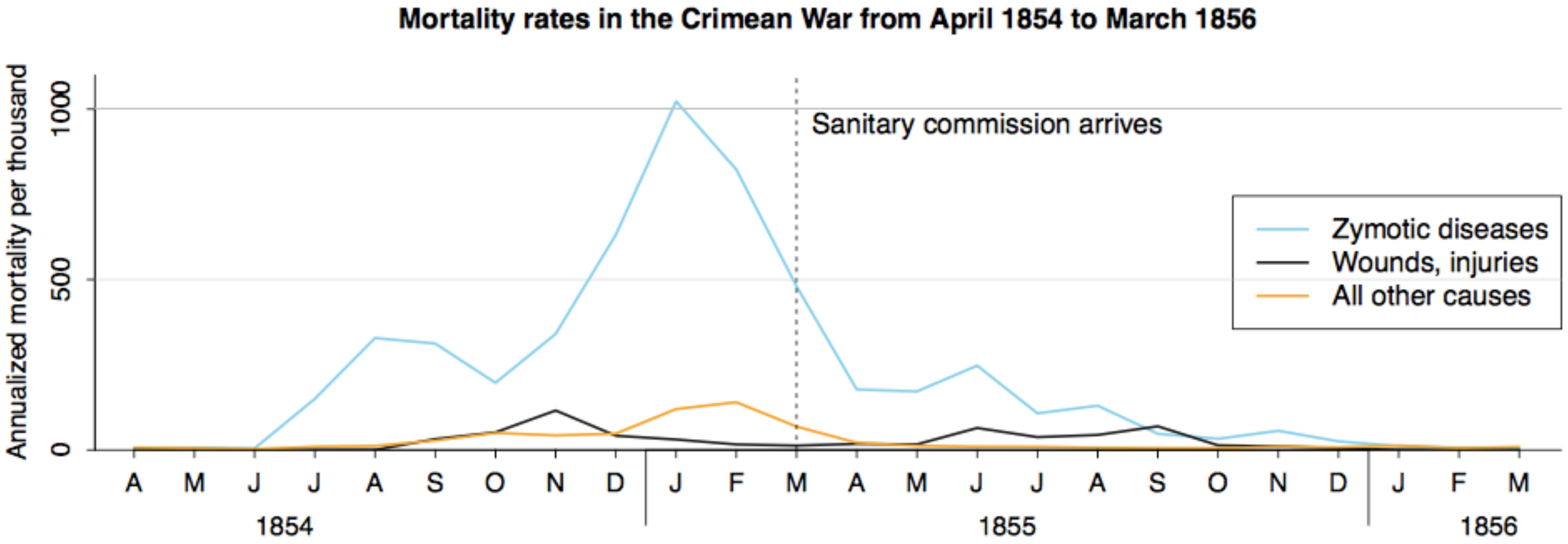


Registered Deaths

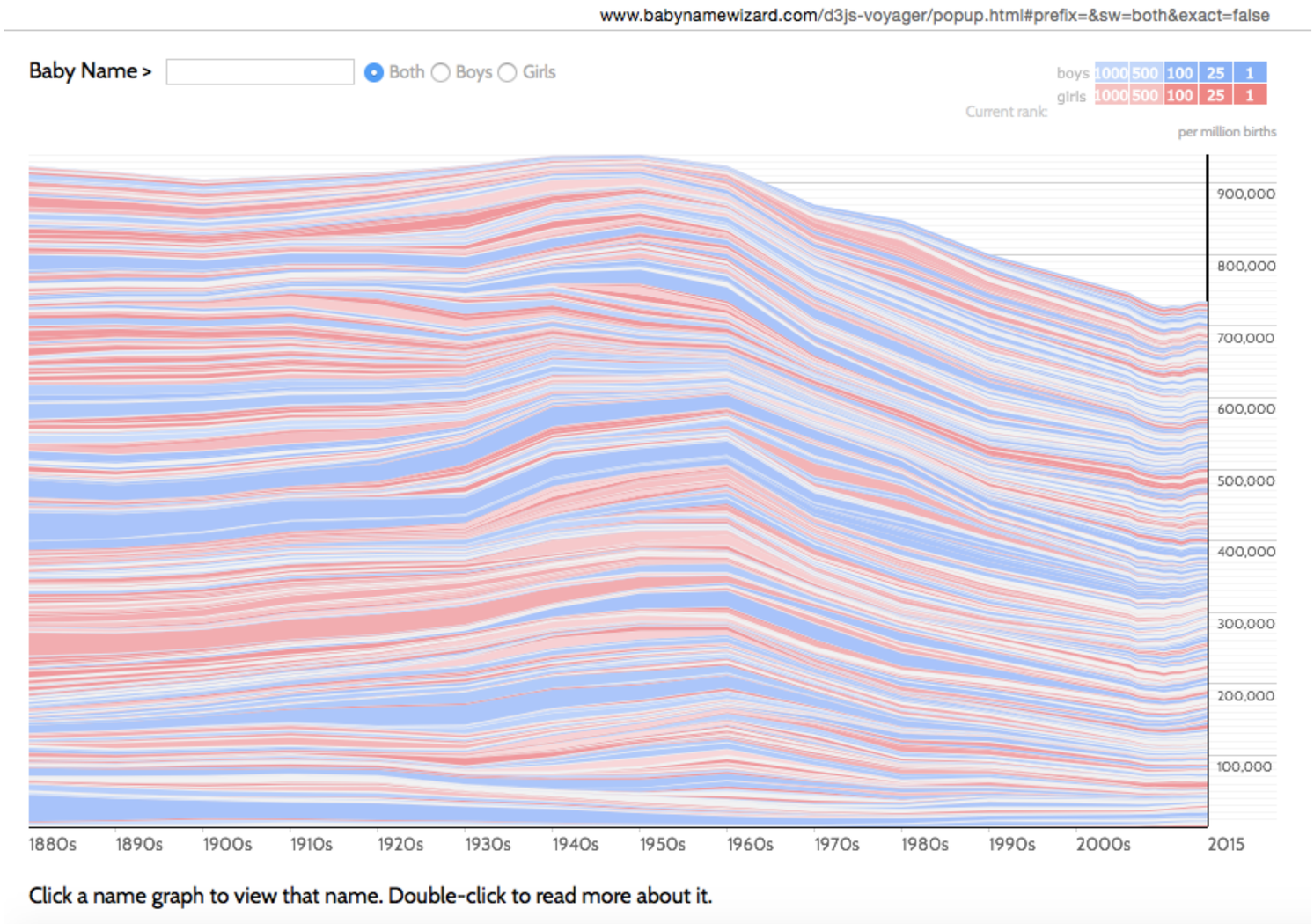
Florence Nightingale's Coxcomb Diagram, 1858



Nightingale's Data, Redrawn



Baby Name Wizard



CODING

Ideal World

RStudio is seeking its next Software Engineer! You'll be joining a team of passionate, talented developers who have a proven track record of producing great software used by hundreds of thousands of data analysts worldwide. We take open-source seriously, and make heavy investments into open-source software in the R community.

You'll be working with a variety of cutting-edge technologies like Go and Angular. **But if you're the kind of developer we're looking for, you'll be able to learn the necessary languages and technologies as you go so we're not worried about which ones you know at the moment.**

https://www.rstudio.com/about/careers/, accessed 1/11/17

More thoughts on coding

- Coding is a tool we use to create data visualizations
- Don't confuse means (tools) with ends (data insights)
- Some tools are better than others
- Success is not determined primarily by the tool

What tools will be learning/using?

- Start with a strong focus on exploratory data analysis w/ R (base, ggplot2)
- Will build coding repertoire through R with Shiny, Plotly, Tableau, intro to web-based
- Later in the semester: focus shifts more to presentation
- Tools for final project are up to you

Tools

https://www.pinterest.com/pin/381257924684737382

Tools ≠ House

https://www.pinterest.com/pin/381257924684737382

http://theberry.com/2011/11/17/my-dream-house-assembly-required-42-photos/

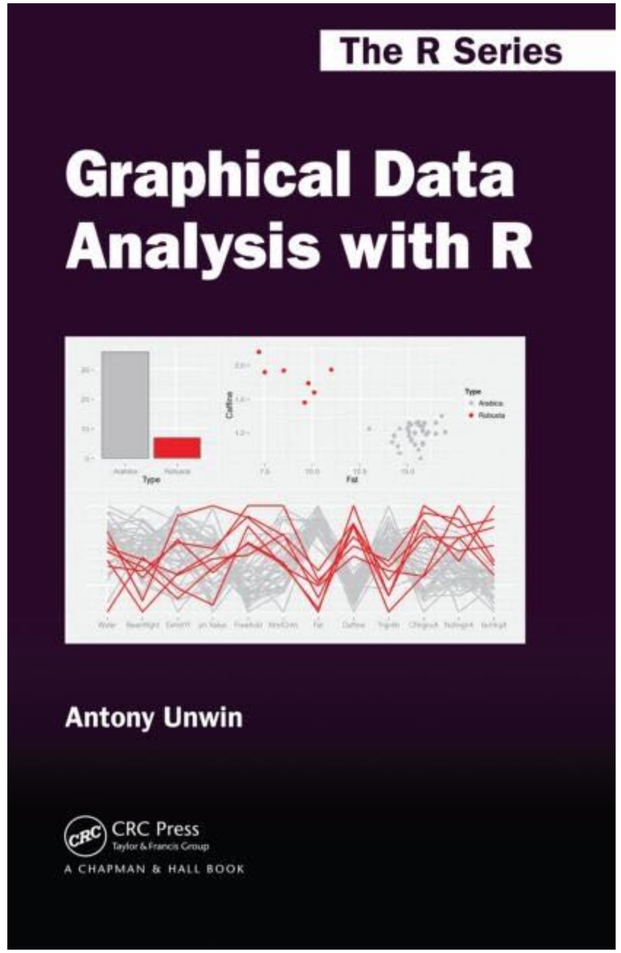
SYLLABUS

Syllabus

(Site for waitlisted students: https://github.com/jtr13/5293)

Required Book

Antony Unwin
2015
Graphical Data Analysis with R
CRC Press
ISBN 978-1498715232



Sources

"John Snow, Cholera Map" https://www1.udel.edu/johnmack/frec682/cholera/

"Data Science Process" diagram: Hadley Wickham and Garrett Grolemund, R for Data Science, 1.1 r4ds.had.co.nz/introduction.html

"Growth of Data Visualization": Hadley Wickham, 2013, "Graphical Criticism: Some Historical Notes", p. 43 www.tandfonline.com/doi/full/10.1080/10618600.2012.761140

"Perception Studies", "Cleveland Dot Plot": Naomi Robbins, 2013, *Creating More Effective Graphs*, Ch. 1., Ch. 3

"Wrong / Misleading Graphs" http://www.mediaite.com/tv/fox-news-airs-seriously-misleading-obamacare-graphic/

"Florence Nightingale's Coxcomb Diagram, 1858" https://understandinguncertainty.org/coxcombs

"Nightingale's Data, Redrawn" Andrew Gelman and Antony Unwin, 2012. "Infovis and Statistical Graphics: Different Goals, Different Looks" http://www.stat.columbia.edu/~gelman/research/published/vis14.pdf