Data visualization

Jay Brophy, Mabel Carabali, Rina Lall

EBOH, McGill University

2024-12-02

Data visualizations

Data Visualizations are...

Data visualizations

Data Visualizations are...

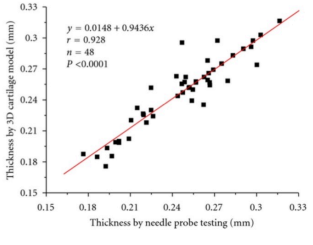
- graphs
- tables
- figures
- maps
- plots
- flowcharts

Data visualizations

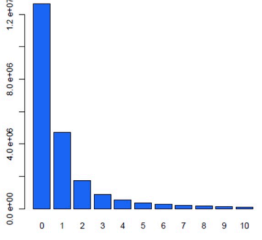
Data Visualizations can...

- help understand basic concepts
- help understand your data
- clarify your story for others
- emphasize a message
- build trust with your audience
- inform / influence their decisions

"The variables are related"

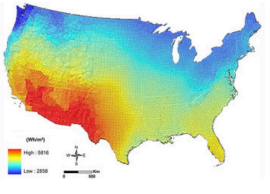


"The distribution is skewed"



"The groups differ in Y"

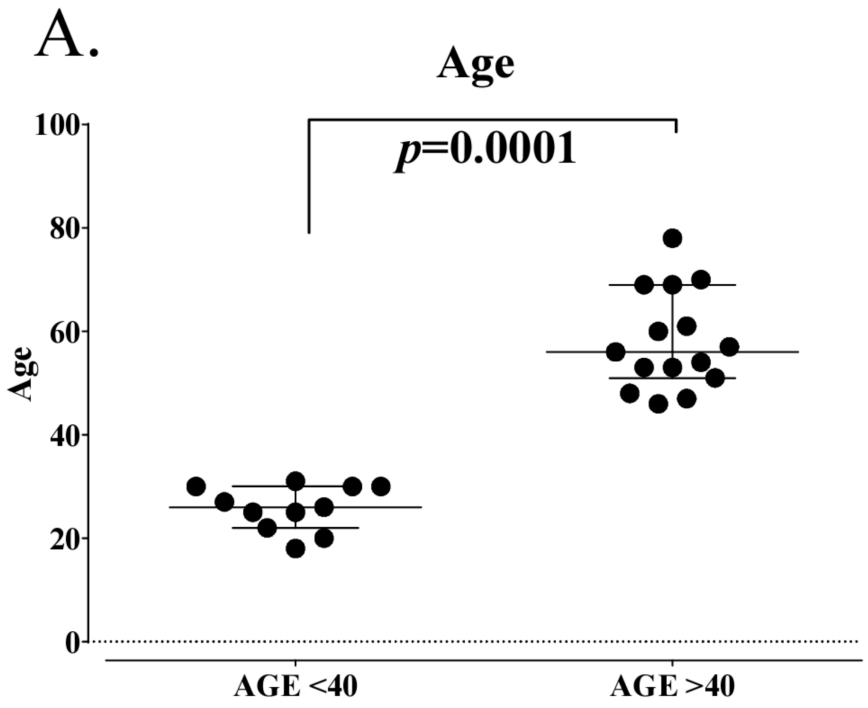
"Affects the Southwest"



Poor data visualizations can do the opposite!

Would you trust these authors?

Even if they come from **National Institutes of Health, Bethesda, MD, USA**



Reference: doi.org/10.5772/62322

Tables are also a form of data visualizations

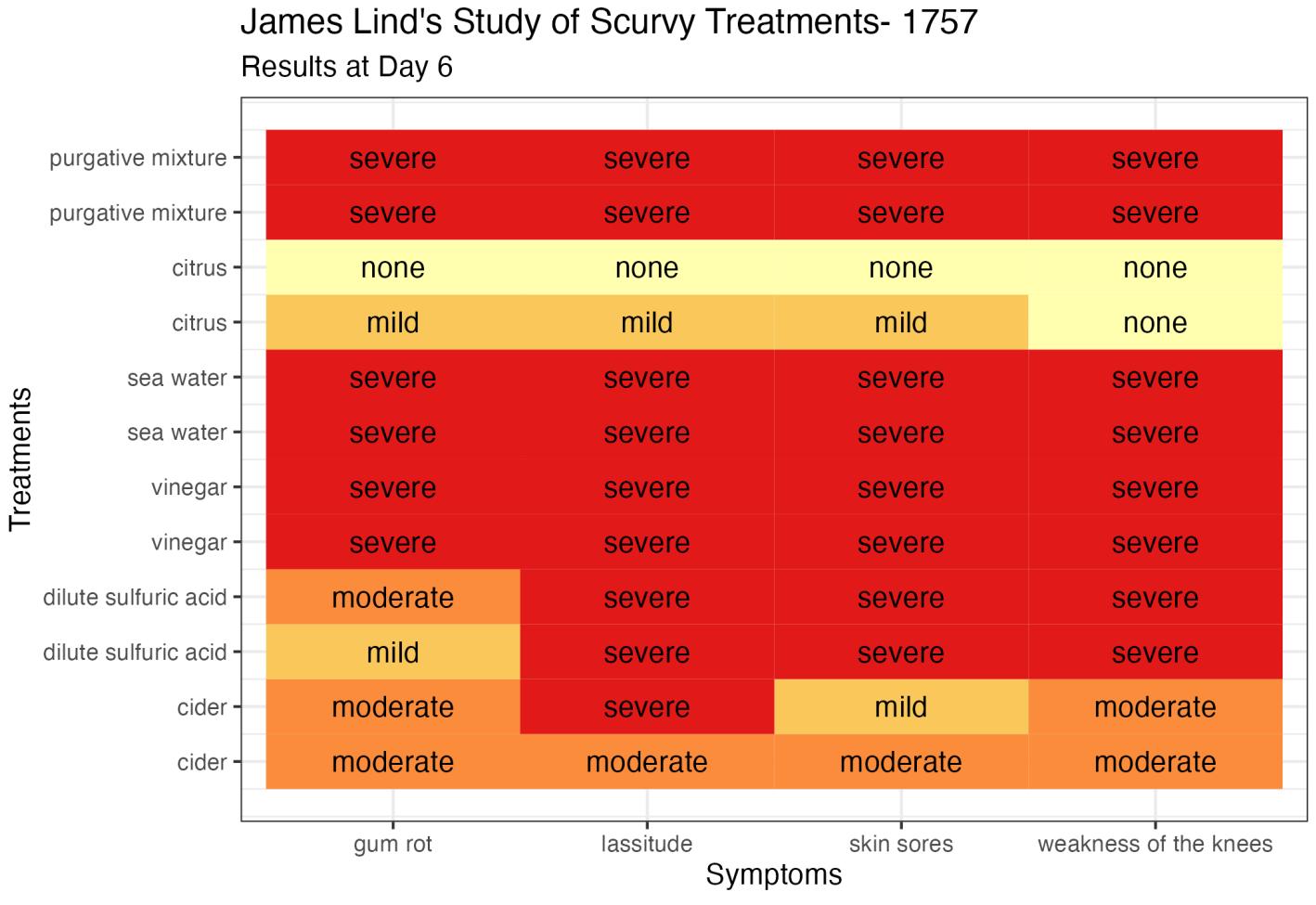
```
scurvy <- medicaldata::scurvy
gt(scurvy)
```

study_id	treatment	dosing_regimen_for_scurvy	gum_rot_d6	skin_sores_d6	weakness_of_the_knees_d6	las
001	cider	1 quart per day	2_moderate	2_moderate	2_moderate	2_
002	cider	1 quart per day	2_moderate	1_mild	2_moderate	3
003	dilute_sulfuric_acid	25 drops of elixir of vitriol, three times a day	1_mild	3_severe	3_severe	3
004	dilute_sulfuric_acid	25 drops of elixir of vitriol, three times a day	2_moderate	3_severe	3_severe	3
005	vinegar	two spoonfuls, three times daily	3_severe	3_severe	3_severe	3
006	vinegar	two spoonfuls, three times daily	3_severe	3_severe	3_severe	3
007	sea_water	half pint daily	3_severe	3_severe	3_severe	3
008	sea_water	half pint daily	3_severe	3_severe	3_severe	3
009	citrus	two lemons and an orange daily	1_mild	1_mild	0_none	

How would you make this table more informative?

For who?

More informative table



data from medicaldata R package

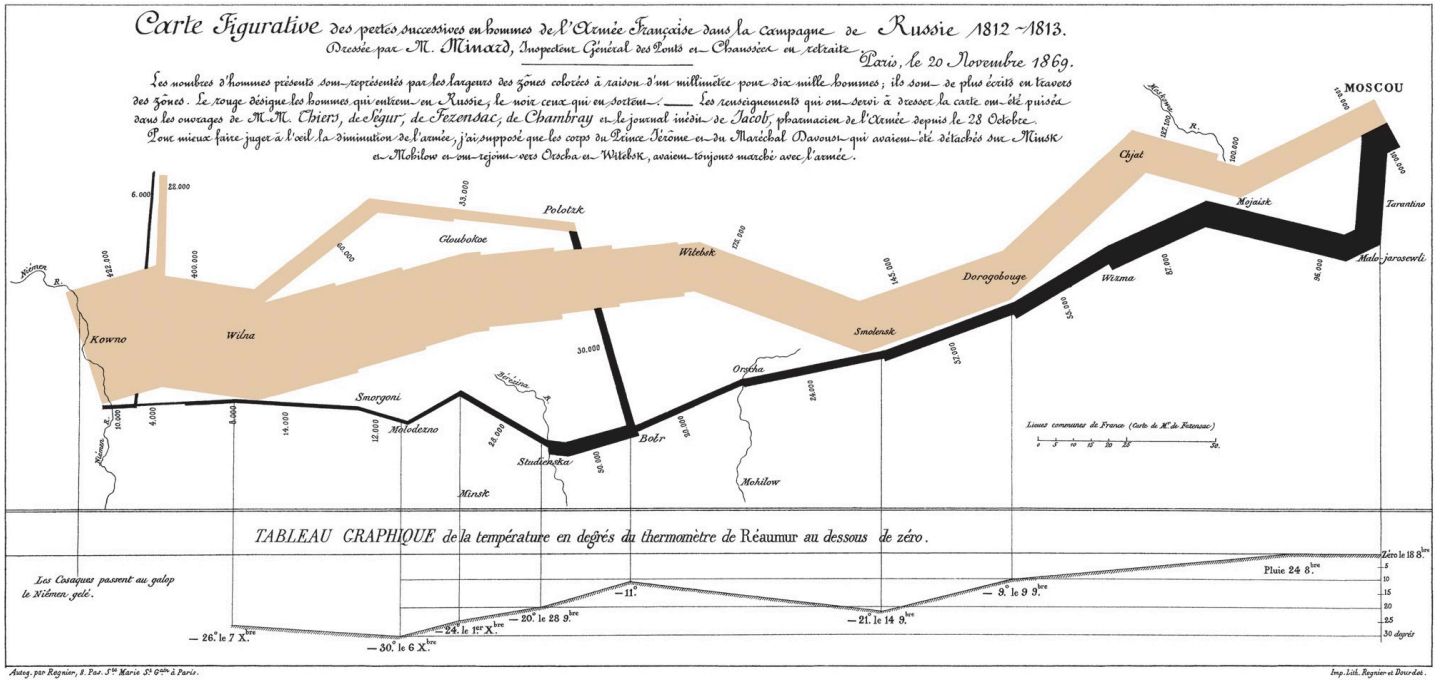
What about non-numeric data?

We can also borrow techniques from other fields...

And find alternative ways to present traditionally tabular data...

Classic visualizations

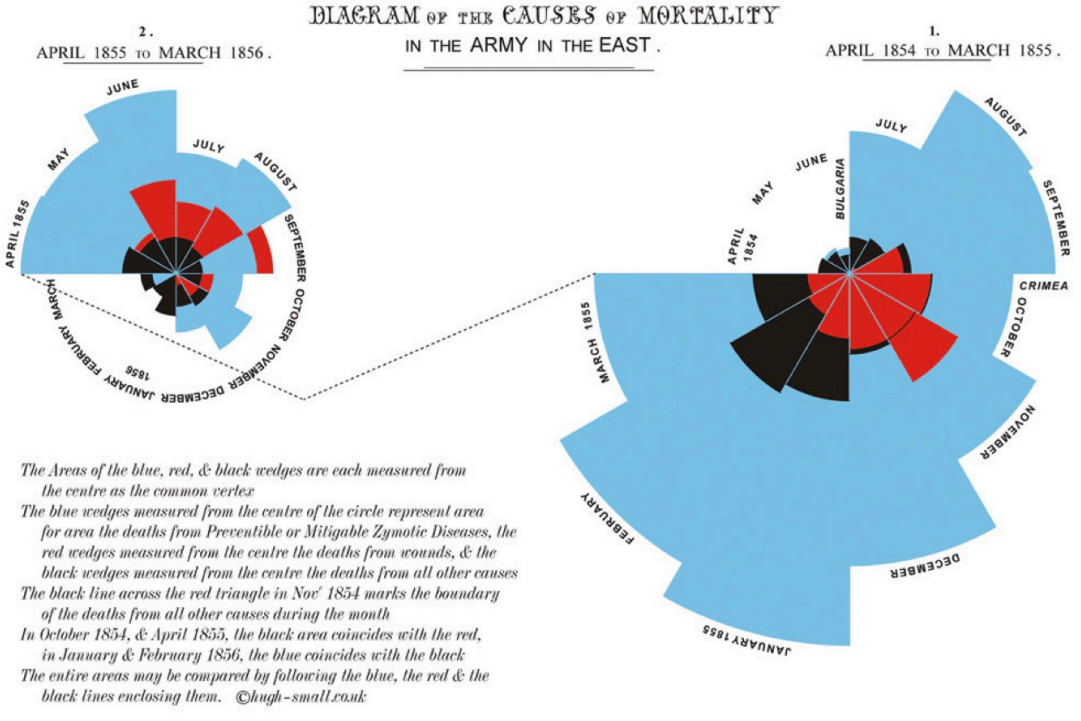
The best statistical graphic ever drawn - Edward Tufte



- 6 types of information: geography, time, temperature, course, direction of the army's movement, # of troops remaining
- "C'est la Bérézina!"

Classic visualizations

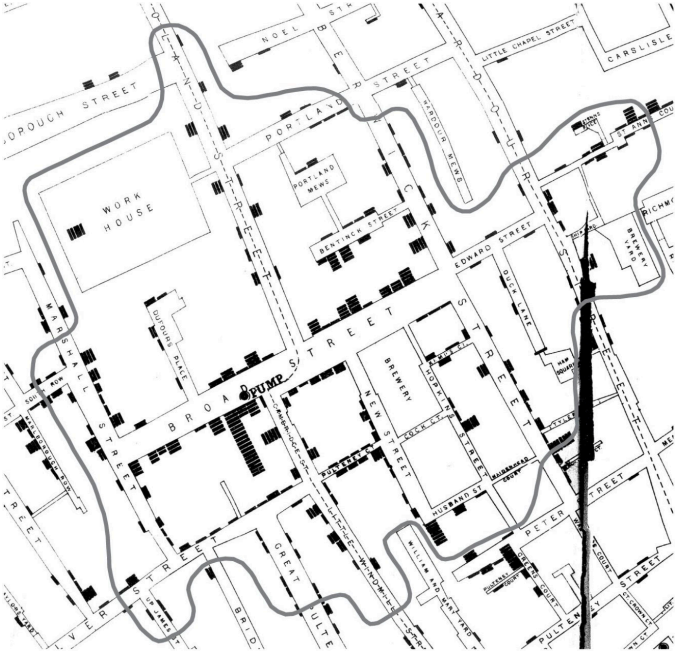
Nightingale's Rose



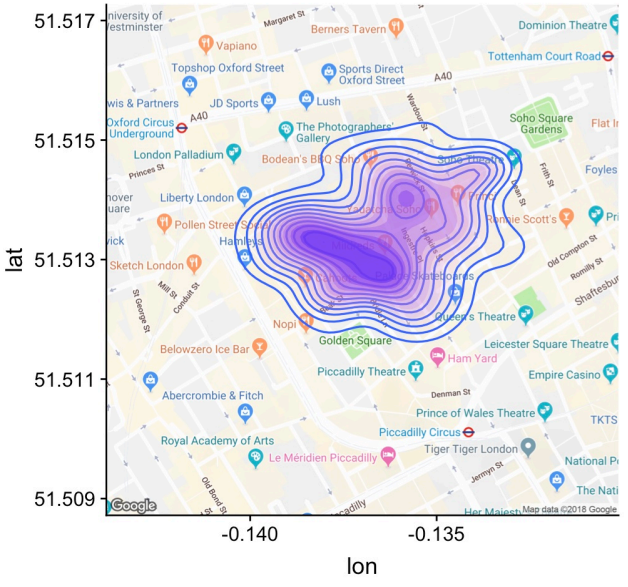
- Although remembered as the mother of modern nursing, Nightingale was an accomplished statistician
- 1st female fellow of the Royal Statistical Society

Another classic visualization

John Snow and the “Ghost Map”



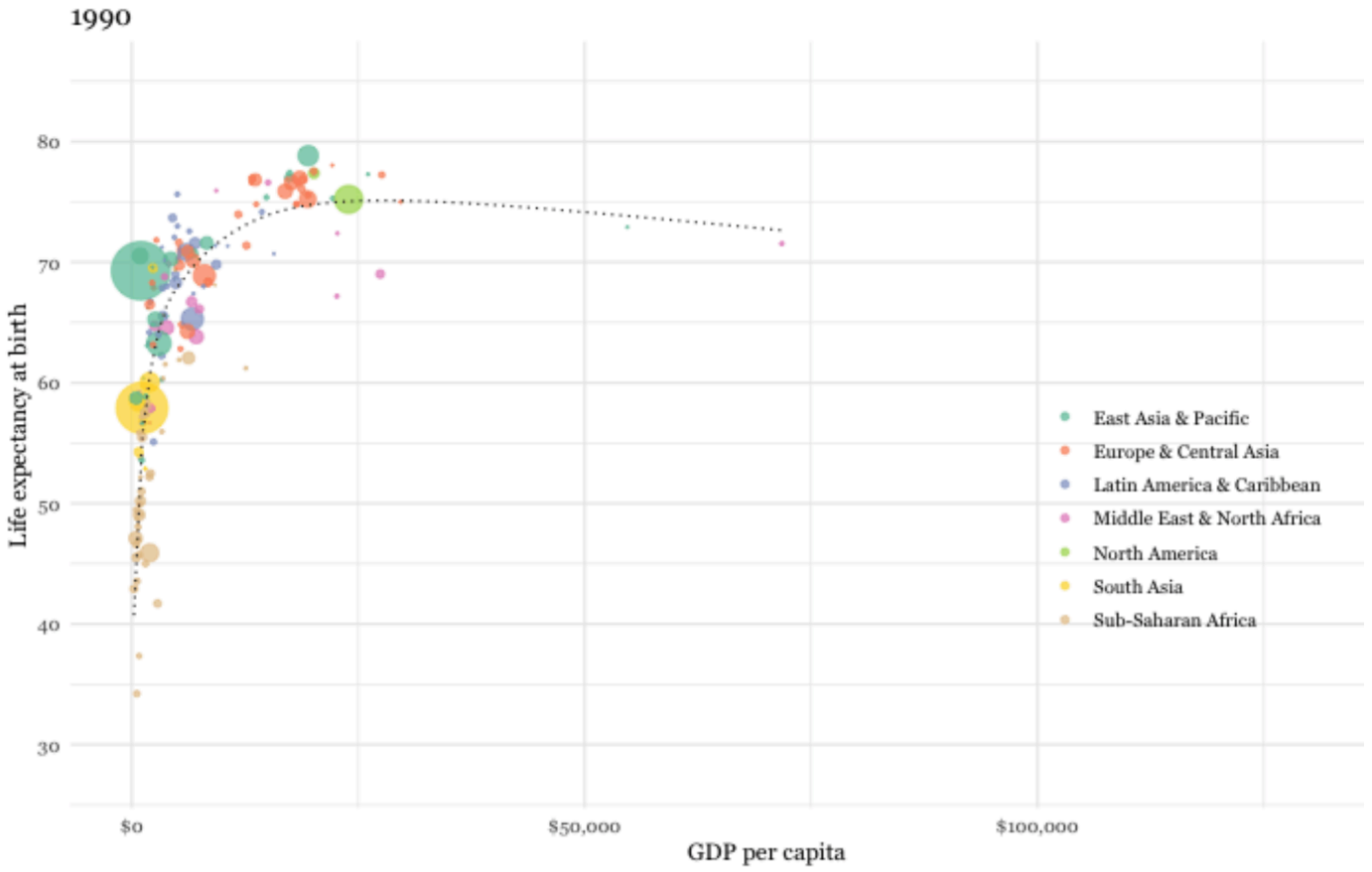
John Snow and the “Ghost Map”
(updated with Goggle maps & R)



Reference: The Ghost Map: The Story of London's Most Terrifying Epidemic--and How It Changed Science, Cities, and the Modern World

Modern Visualizations

Modern approach - watch Hans Rosling



Modern Visualizations

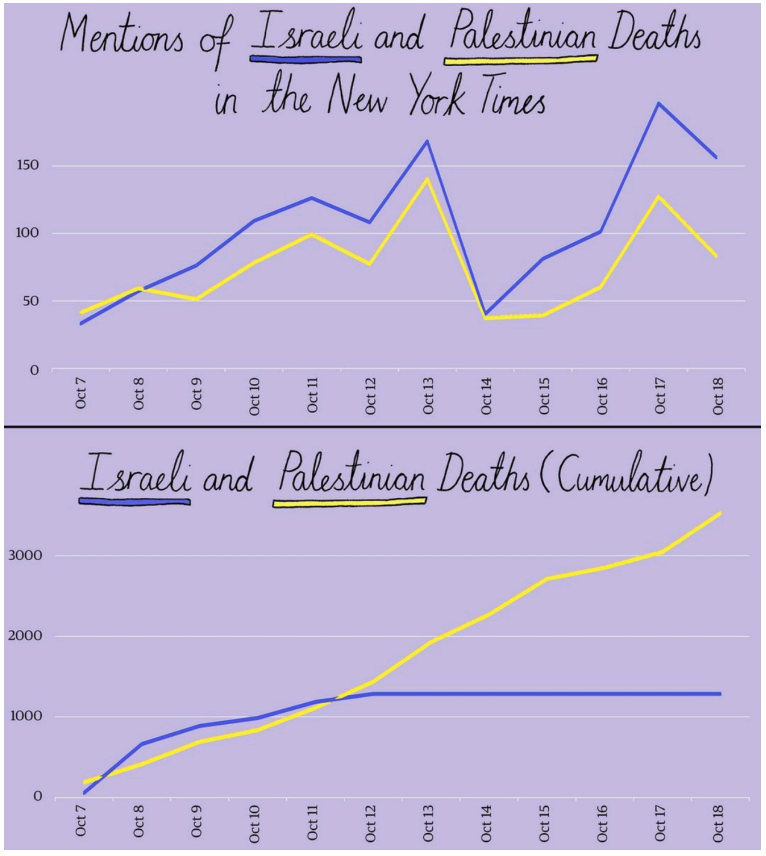
Mona Chalabi Instagram

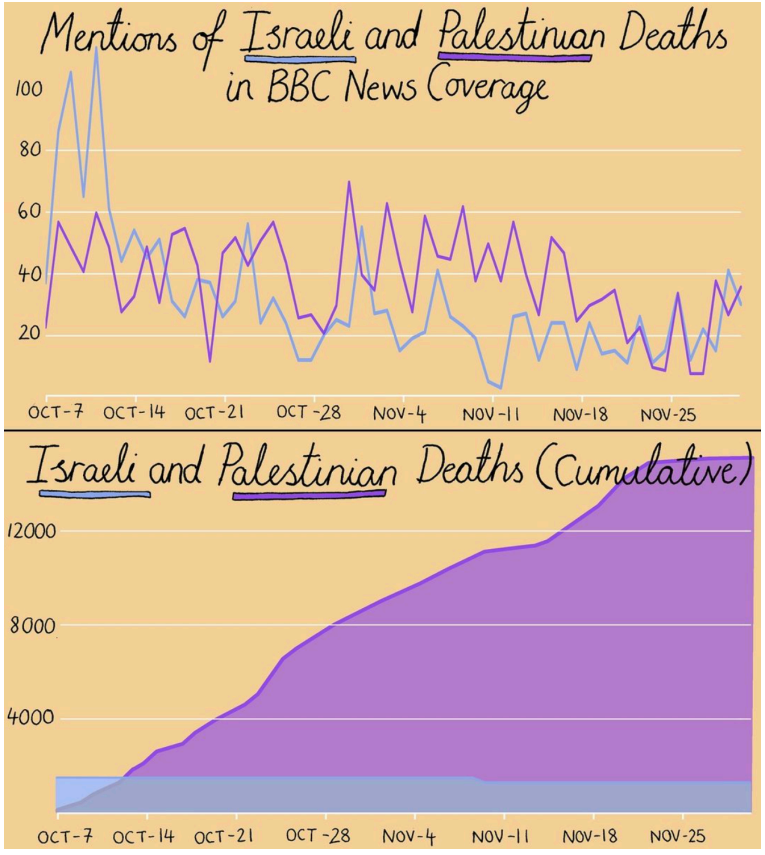
Mona Chalabi is an award-winning writer and illustrator.

Her work has earned her a Pulitzer Prize, a fellowship at the British Science Association, an Emmy nomination and recognition from the Royal Statistical Society. Her writing and illustrations have been featured in The New York Times, The New Yorker and The Guardian where she is currently the data editor.

Reference: https://monachalabi.com/

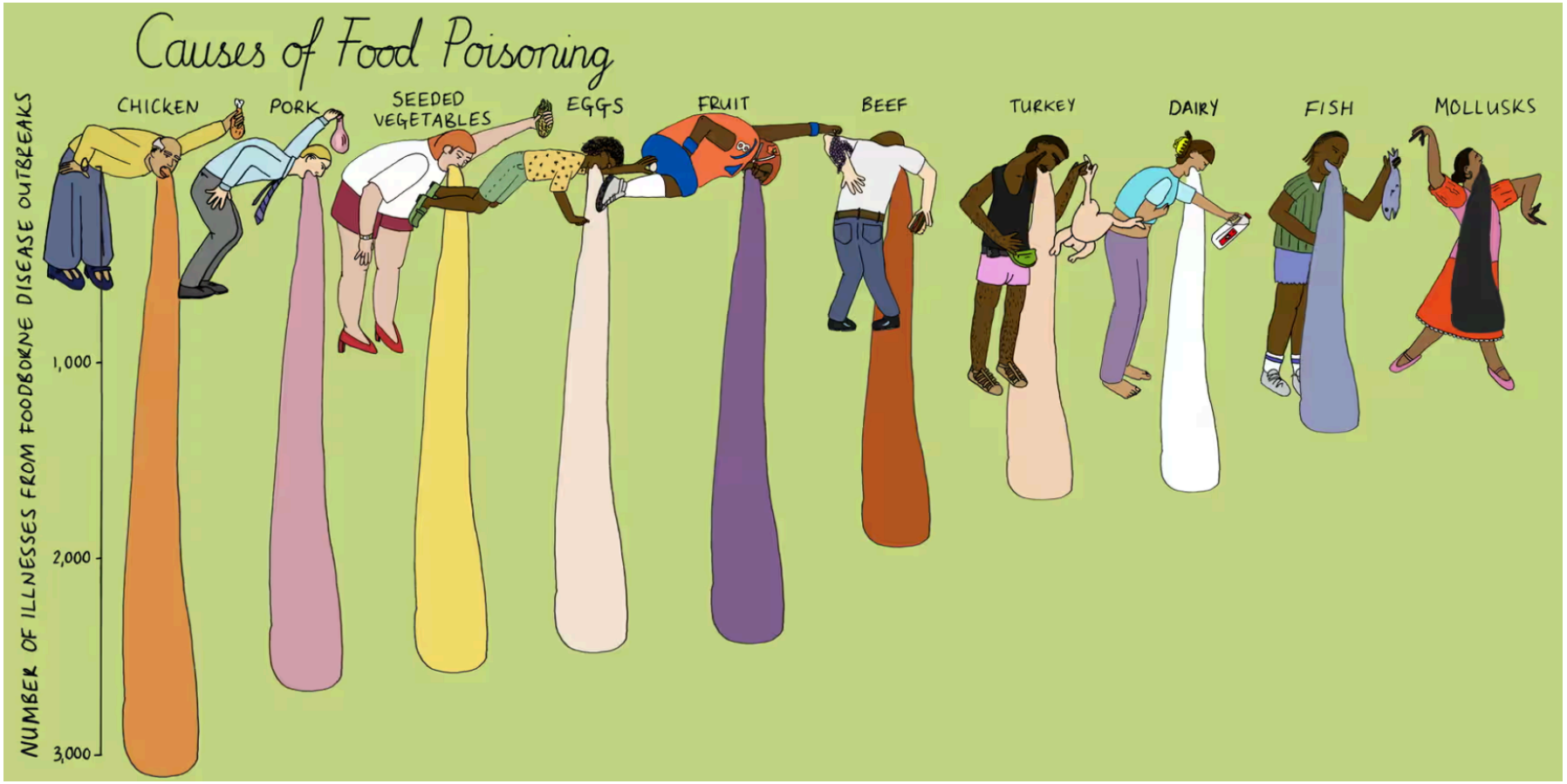
Modern Visualizations





Reference: Instagram

Modern Visualizations



Reference: The Guardian

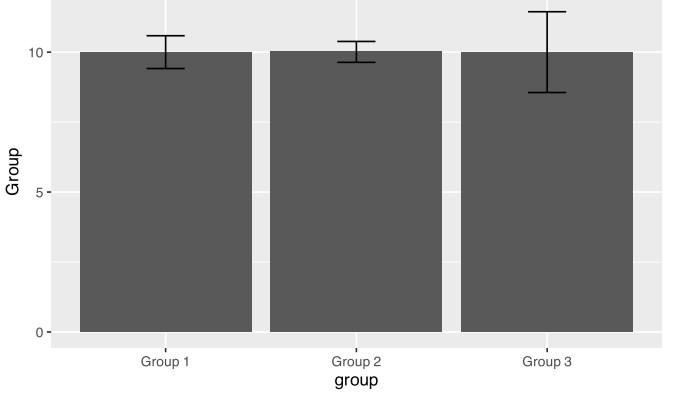
Uninspired visualization (barplot)

Generate some data

```
set.seed(2021)
data <- tibble(
    group = factor(c(rep("Group 1", 100), r
    value = c(## Group 1
        seq(0, 20, length.out = 100),
        ## Group 2
        c(rep(0, 5), rnorm(30, 2, .1), rnor
        ## Group 3
        rep(seq(0, 20, length.out = 5), 5))
) %>%
    rowwise() %>%
    mutate(value = if_else(group == "Group
```

Barplots are often used to display and summarize the data but are uninspired

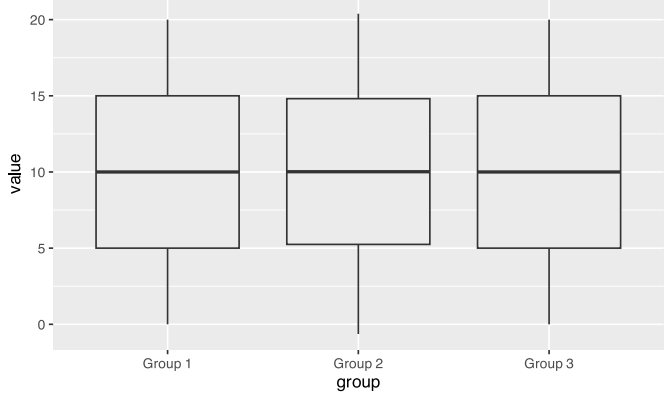
```
ggplot(data, aes(x=group, y=value)) + geom_
      ylab("Average value") + scale_y_conti
```



Semi-inspired visualization (boxplot)

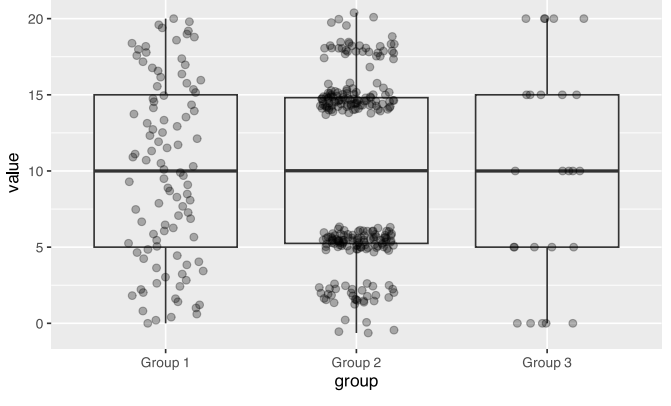
Boxplots are a marginal improvement over barplots in that they show the median, IQR and outliers but are still uninspired

```
gg1 <- ggplot(data, aes(x = group, y = valu
gg1
```



Big improvement adds the data points

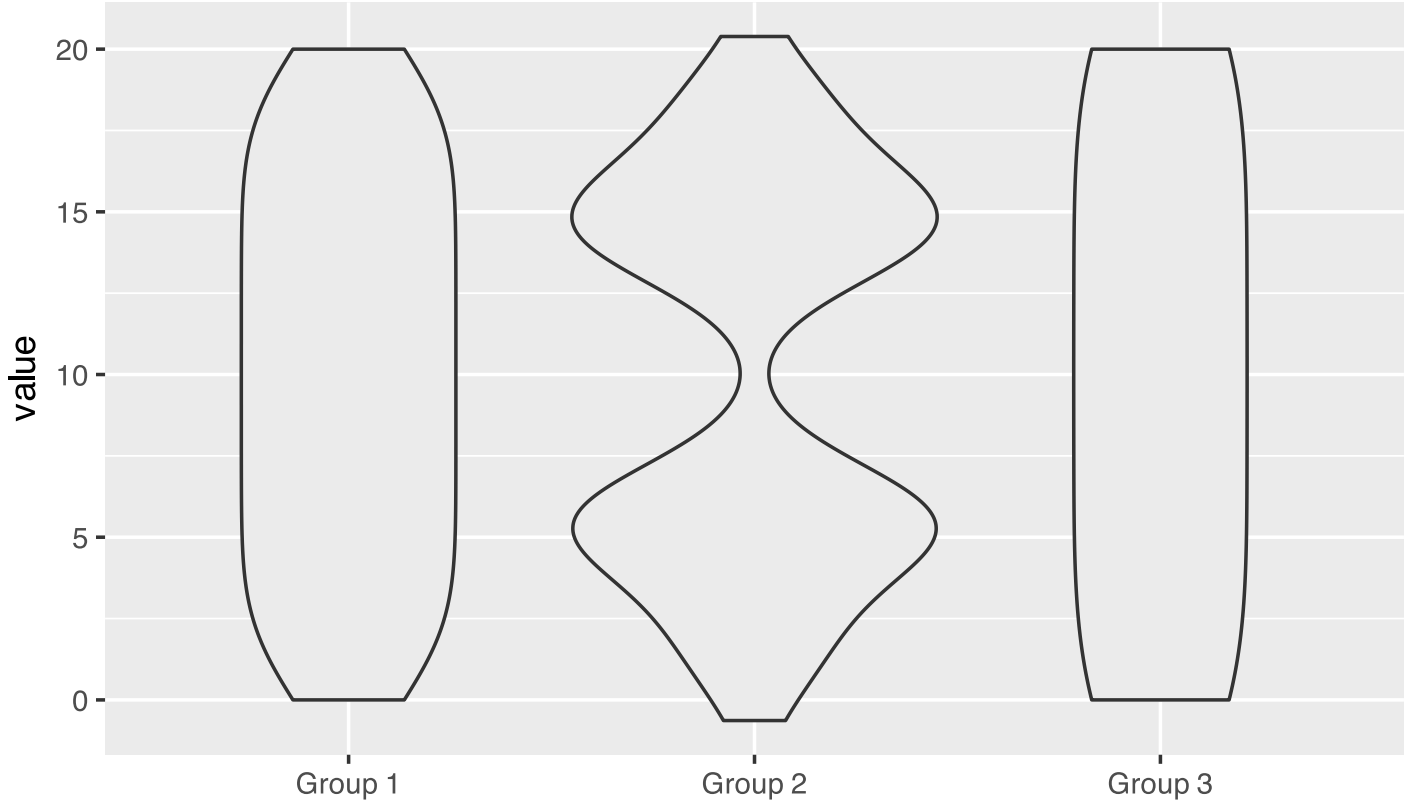
```
gg1 + geom_point(## draw bigger points
    size = 2, ## add some transparency
    alpha = .3, ## add some jittering
    position = position_jitter(## control r
      seed = 1, width = .2))
```



Violin plots

Violin plots enable one to see the distribution of the data

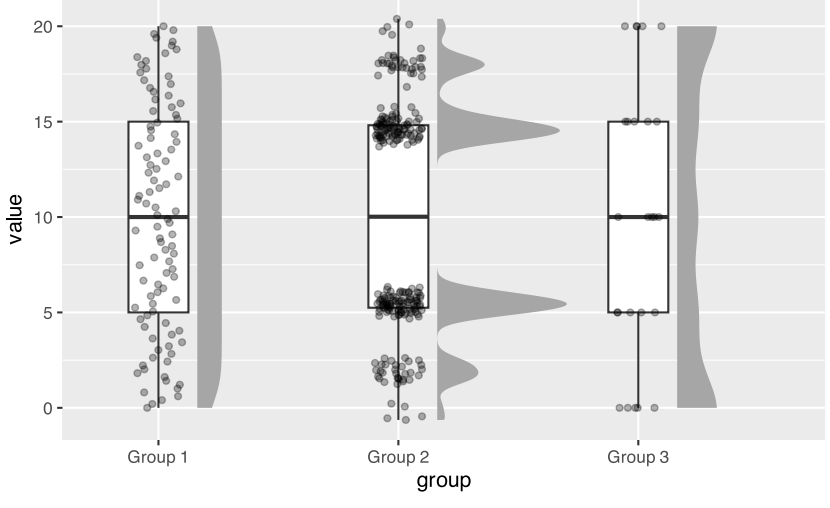
```
ggplot(data, aes(x = group, y = value)) +
  geom_violin(fill = "grey92")
```



Inspirational visualization - Rain clouds

Rainclouds build on the combined violin / boxplot by adding the raw data as well from Cederic Scherer's amazing blog

```
gg4 <- ggplot(data, aes(x = group, y = valu
  ggdist::stat_halfeye(
    adjust = .5, width = .6, .width = 0,
    justification = -.3, point_colour = NA)
  geom_boxplot(
    width = .25, outlier.shape = NA) +
  geom_point(
    size = 1.3,
    alpha = .3,
    position = position_jitter(seed = 1, wi
  ) +
  coord_cartesian(xlim = c(1.2, NA), clip =
```



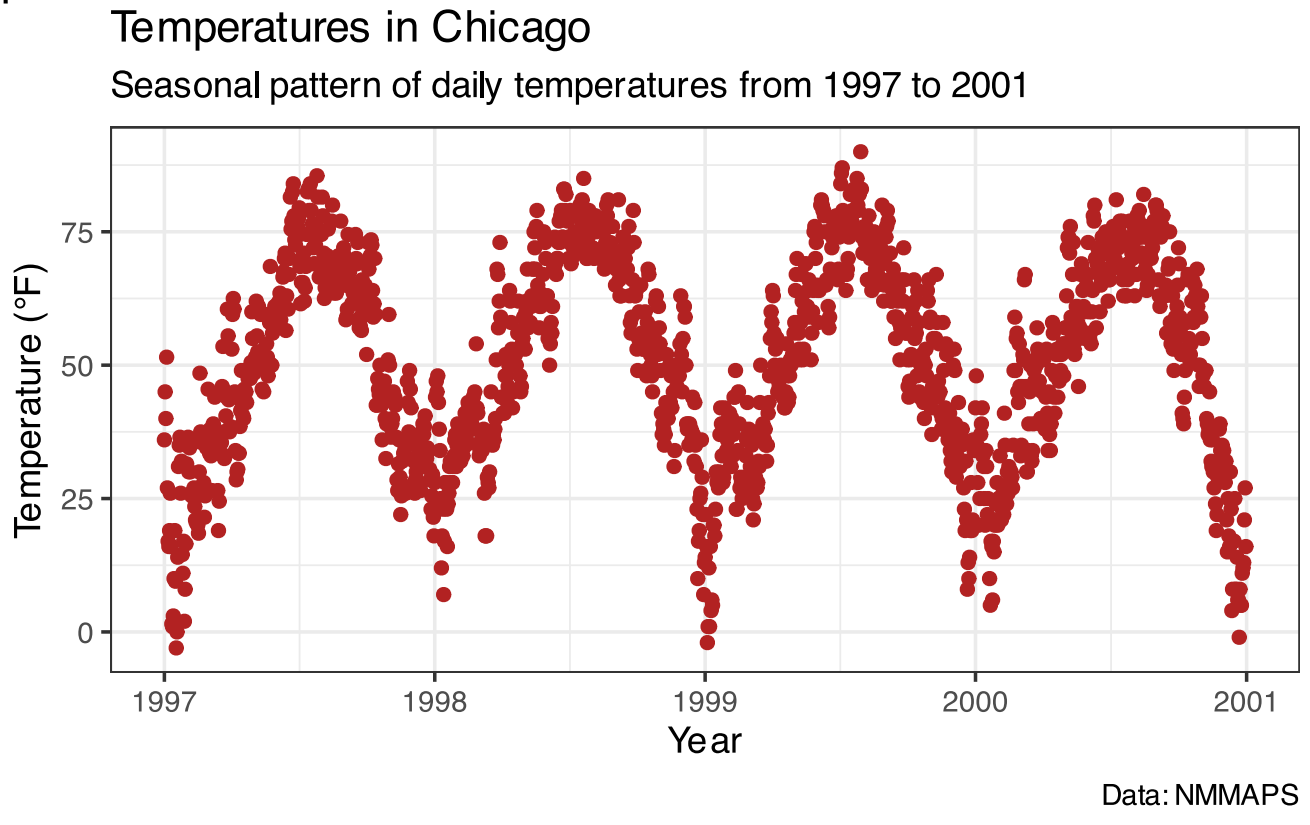
Quick tutorial for ggplot2

There are many terrific sources for tutorials and examples for `ggplot`
One example is found here

```
chic <- readr::read_csv("chic.csv", show_col_types = FALSE) # read data set - weather in # C
#data(chic, package = "data.704")
g <- ggplot(chic, aes(x = date, y = temp)) + #set data and axes
  geom_point(color = "firebrick") + # add color, could also change "shape" and "size"
  labs(x = "Year", y = "Temperature (°F)",
       title = "Temperatures in Chicago",
       subtitle = "Seasonal pattern of daily temperatures from 1997 to 2001",
       caption = "Data: NMMAPS",
       tag = "Fig. 1") + # title, labels, etc
  theme(axis.text.x = element_text(angle = 50, vjust = 1, hjust = 1, size = 12)) + # rotate
  theme_bw() # change the  theme (background, etc)
```

Quick tutorial for ggplot2

Fig. 1

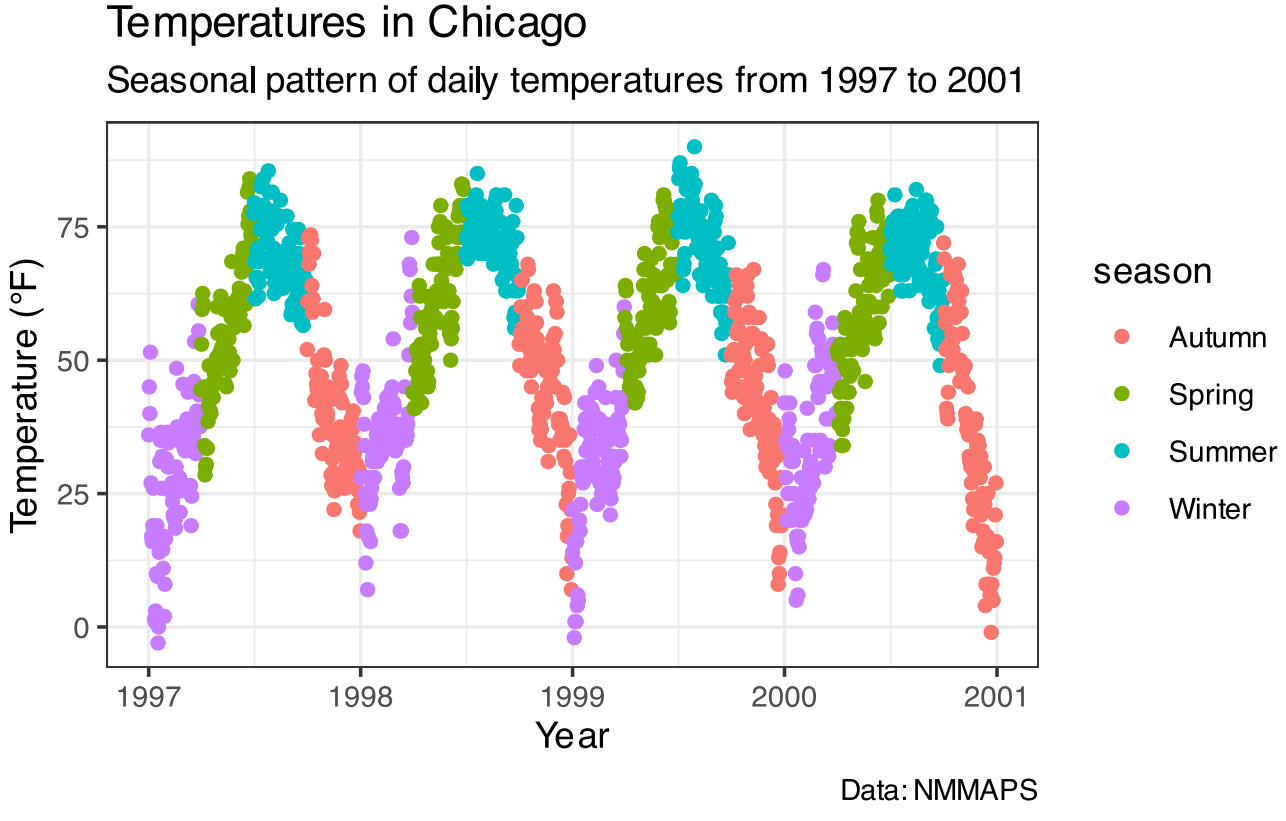


Quick tutorial for ggplot2

Can code the plot based on season, i.e. map the variable season to the aesthetic color

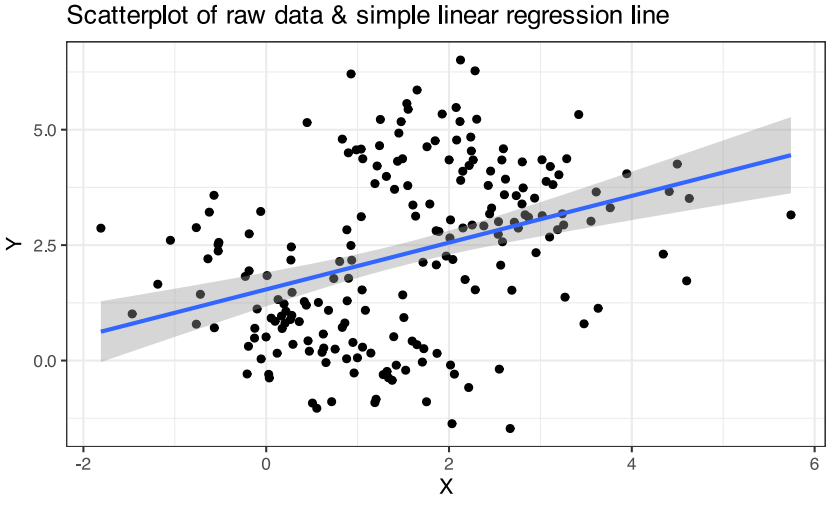
```
ggplot(chic, aes(x = date, y = temp, color = season)) +
```

Fig. 1



Graphical understanding of confounding

```r
#Generate some data
set.seed(123)
df <- data.frame(W = as.integer((1:200>100)
  mutate(X = .5+2*W + rnorm(200)) %>%
  mutate(Y = -.5*X + 4*W + 1 + rnorm(200),t
  group_by(W) %>%
  mutate(mean_X=mean(X),mean_Y=mean(Y)) %>%
  ungroup()

g <- ggplot(df, aes(x=X,y=Y)) + geom_point(
  geom_smooth(method=lm) + ggtitle("Scatter
  theme_bw()
```



But the dataset includes another variable **W** that affects **X** and **Y**

Multiple regression

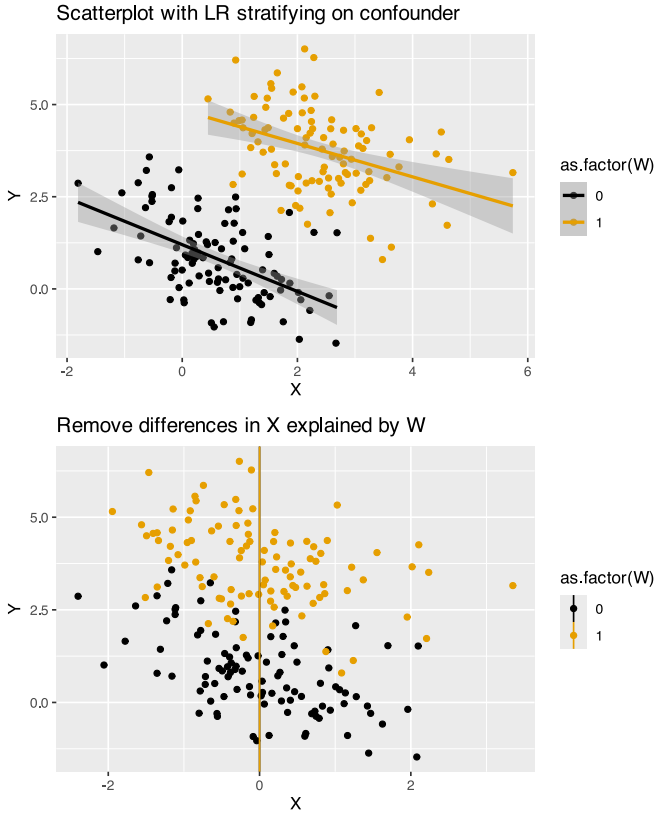
```
mod <- lm(Y ~ X + W, data = df)
summary(mod)
```

```
##
## Call:
## lm(formula = Y ~ X + W, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.38799 -0.68148 -0.06722  0.69909  2.59742
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.14316    0.10920  10.469  < 2e-16 ***
## X           -0.53844    0.07537  -7.144 1.71e-11 ***
## W            3.91258    0.19579  19.984  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9972 on 197 degrees of freedom
## Multiple R-squared:  0.7117,    Adjusted R-squared:  0.7088
## F-statistic: 243.2 on 2 and 197 DF,  p-value: < 2.2e-16
```

Notice slope of X parameter is now negative, was positive in previous graph

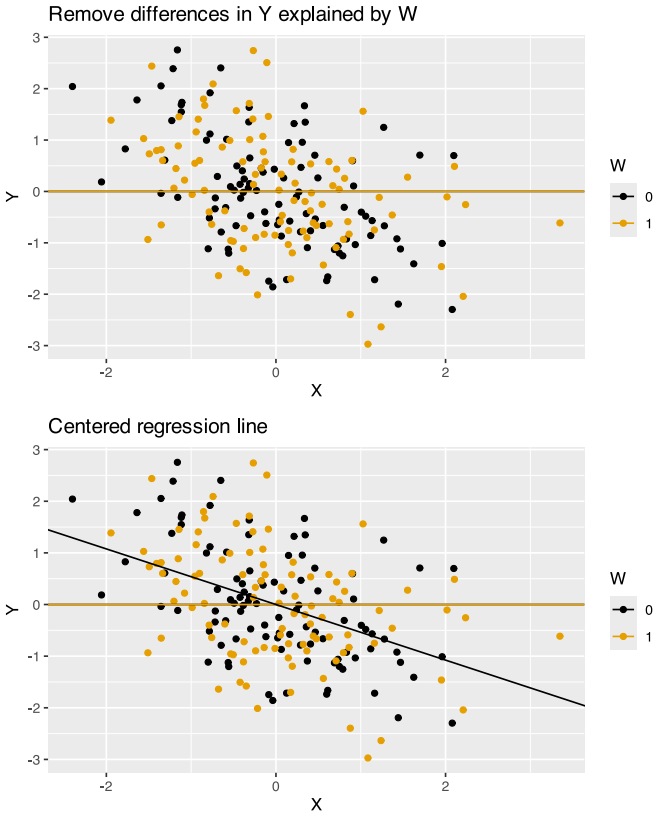
Graphical explanation of confounding

```
# basic plot x vs y + linear regression + s
g1 <- ggplot(df, aes(x=X, y=Y, color=as.fac
  geom_point() +
  ggthemes::scale_color_colorblind() +
  geom_smooth(method=lm) +
  ggtitle("Scatterplot with LR stratifying

#Remove differences in X explained by W
g2 <- df %>% mutate(X = X - mean_X,mean_X=0
  ggplot(aes(y=Y,x=X,color=as.factor(W))) +
  geom_point() +
  geom_vline(aes(xintercept=mean_X,color=as
  ggthemes::scale_color_colorblind() +
  ggtitle("Remove differences in X explaine
```



Graphical explanation of confounding

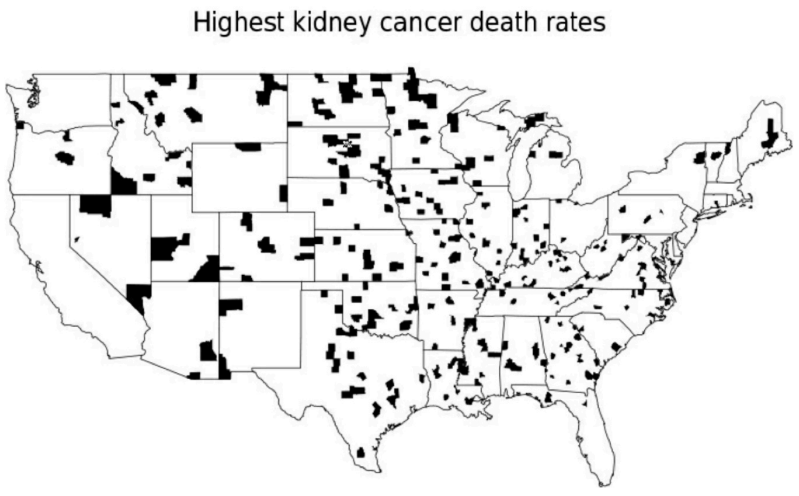
```
#Remove differences in Y explained by W
g3 <- df %>% mutate(X = X - mean_X,Y = Y -
  ggplot(aes(y=Y,x=X,color=as.factor(W))) +
  geom_point() +
  geom_hline(aes(yintercept=mean_Y,color=as
  ggthemes::scale_color_colorblind() +
  guides(color=guide_legend(title="W")) +
  ggtitle("Remove differences in Y explaine

#Centered regression line
g4 <- df %>% mutate(X = X - mean_X,Y = Y -
  ggplot(aes(y=Y,x=X,color=as.factor(W))) +
  geom_point() +
  geom_hline(aes(yintercept=mean_Y,color=as
  ggthemes::scale_color_colorblind() +
  guides(color=guide_legend(title="W")) +
  ggtitle("Centered regression line") +
  geom_abline(intercept = 0, slope = mod$co
```



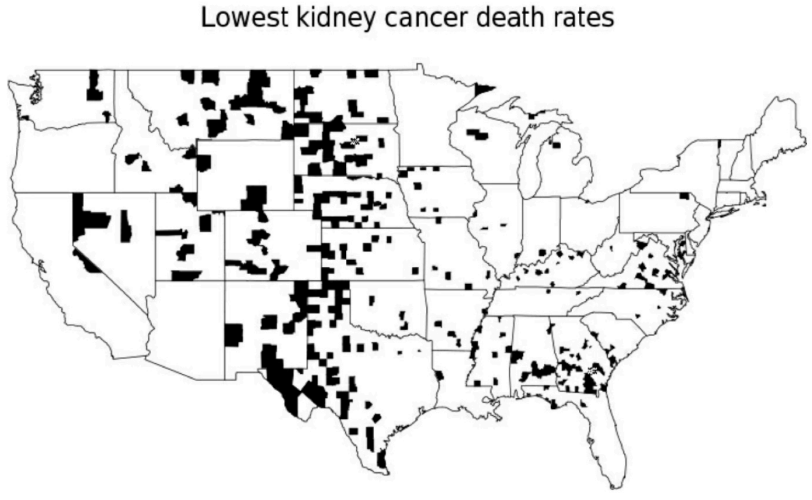
Slope of this graph (-0.05) = parameter estimate of multiple LR

Don't get fooled

US counties with the highest 10% age-standardized death rates for kidney cancer, 1980–1989
Hypotheses?



US counties with the lowest 10% age-standardized death rates for kidney cancer, 1980–1989
Hypotheses?



Reference: Gelman, Andrew. Teaching Statistics (p. 14). OUP Oxford. Kindle Edition

Variability of small numbers!

• In a county with 100 people: if 1 kidney cancer death in 1980s, -> 1 / 1000 per decade, among the highest rates
• If no kidney cancer deaths, rate will be lowest
• Observed rates for smaller counties are much more variable, even if the true cancer probability in these counties is nothing special, probably random fluctuation
• If a large county has a very high rate, more likely a real phenomenon

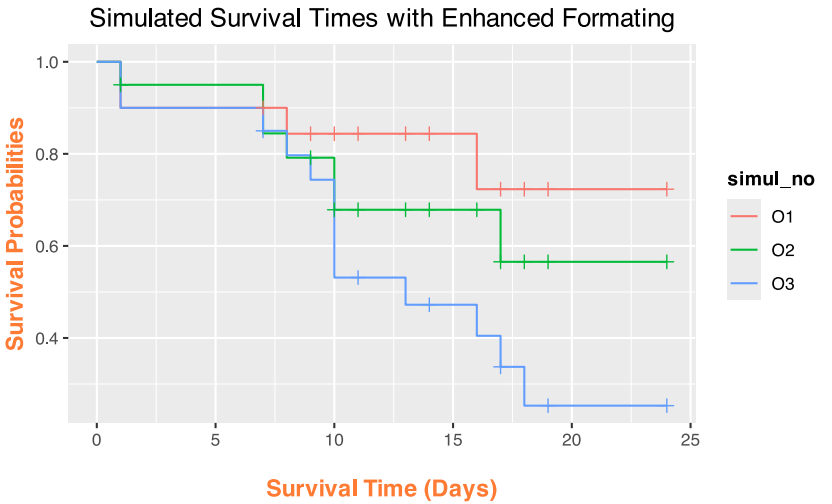
One Solution

Bayes-estimated county rate - a weighted average of (a) the observed rate in the county and (b) the national average rate.
Weights proportional to the population of the county
Counties with extremely small population the rate is shrunk virtually to the national average
Counties with moderate population the rate is shrunk part way toward the national average
Large counties the adjusted rate is essentially equivalent to the observed rate

Better understanding data - example 1

Example Table

ID	O1	O2	O3	t	nb
1	0	0	0	17	1
2	0	0	1	9	1
3	1	1	1	1	1
4	0	0	1	10	1
5	0	0	1	18	1
6	0	0	0	14	1
7	0	0	1	13	1
8	0	1	1	7	1
9	0	0	0	19	1
10	1	0	1	16	1
11	0	0	0	11	1
12	0	0	0	24	1



Better understanding data - example 2

Time for a little Bayes

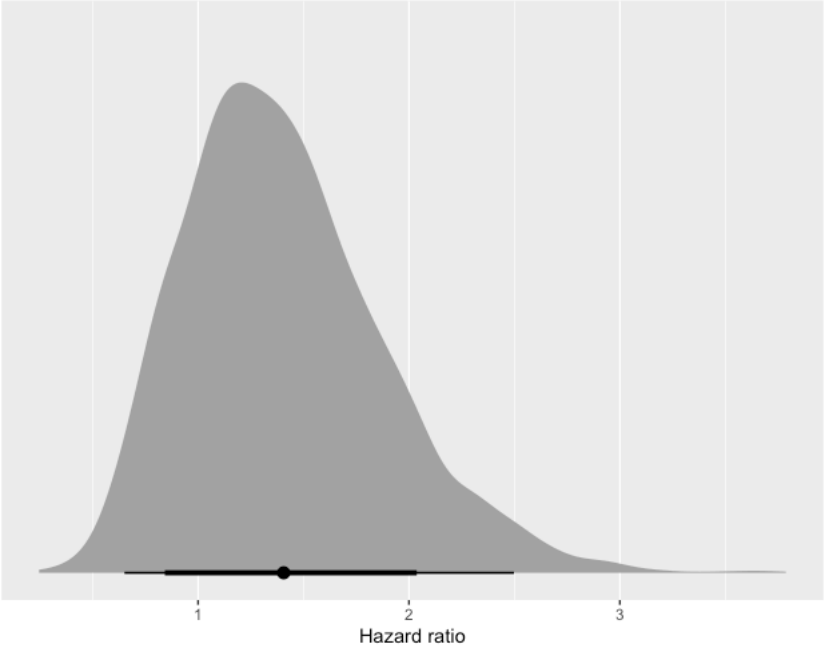
```
summary(mod)
```

```
##  Family: cox
##   Links: mu = log
## Formula: T1 | cens(delta1) ~ agvhd
##    Data: td_dat (Number of observations: 163)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     1.29      0.20     0.90     1.67 1.00     2237     2323
## agvhd         0.28      0.36    -0.47     0.94 1.00     2688     2462
##
## Draws were sampled using sample(hmc). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

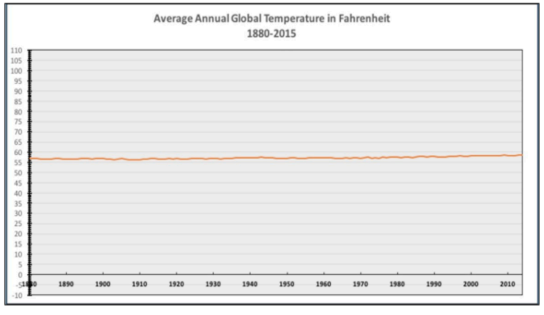
Taking exponentials HR = 1.33, 95% CrI 0.66 - 2.61 compares quite closely to frequentist vales(1.40, 95%CI 0.81 - 2.43)

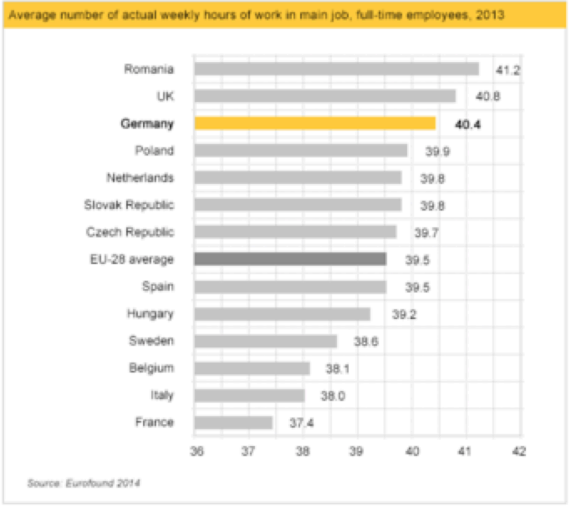
This has used the default vague prior to see the default use `prior_summary`

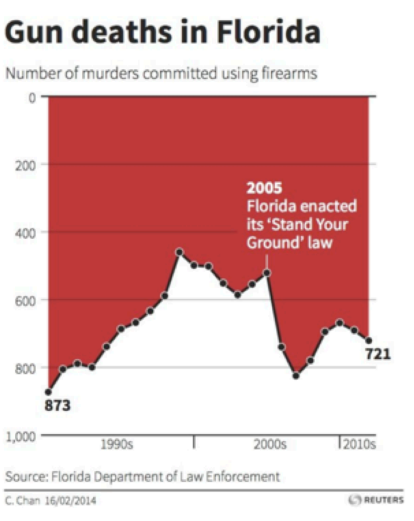
To use informative prior add `prior=c(set_prior())` to the `brm` function



Deceptive graphs

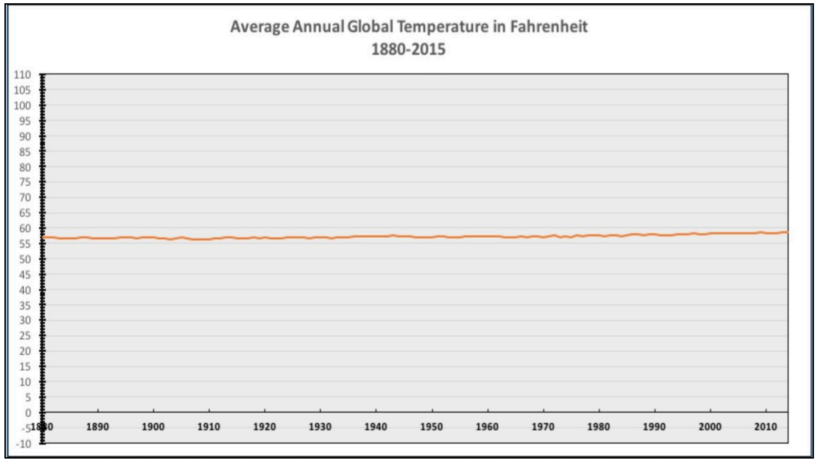




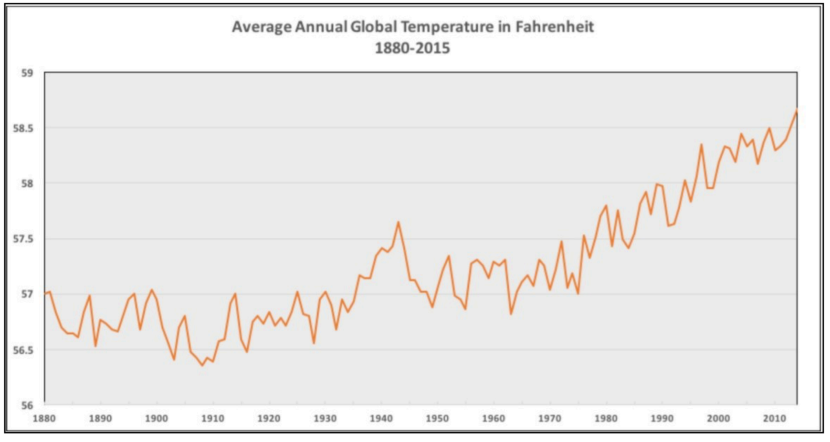


Deceptive graphs #1

Original plot



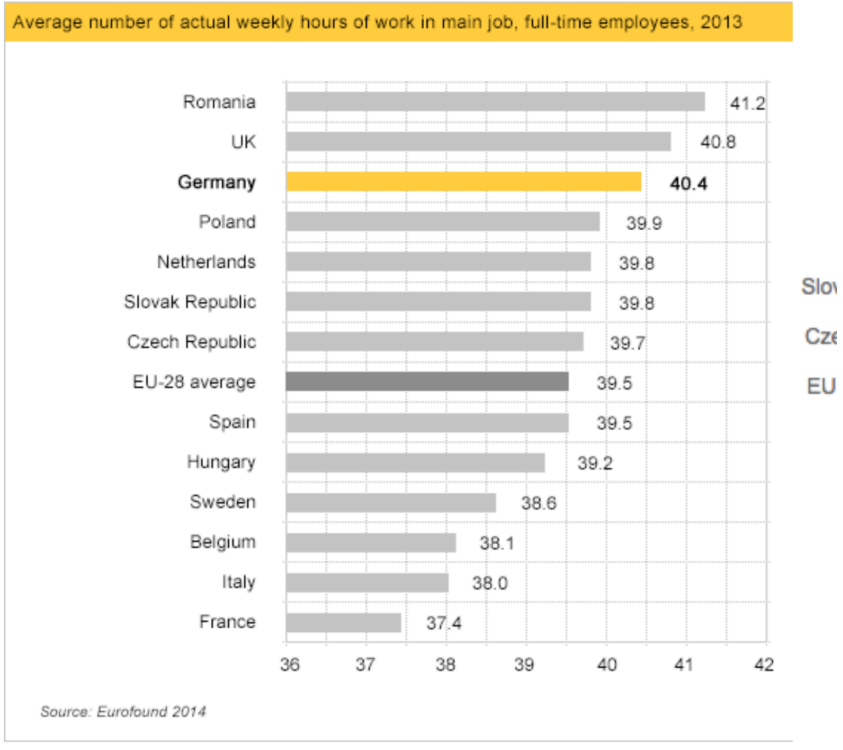
Better plot



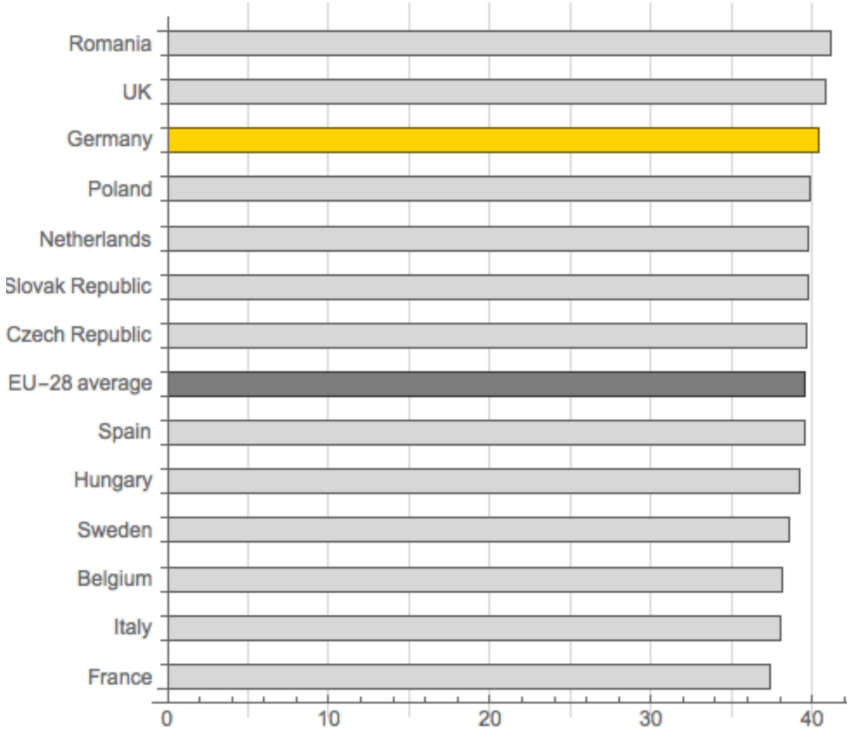
Use the right axis scale for proper interpretations

Deceptive graphs #2

Original plot



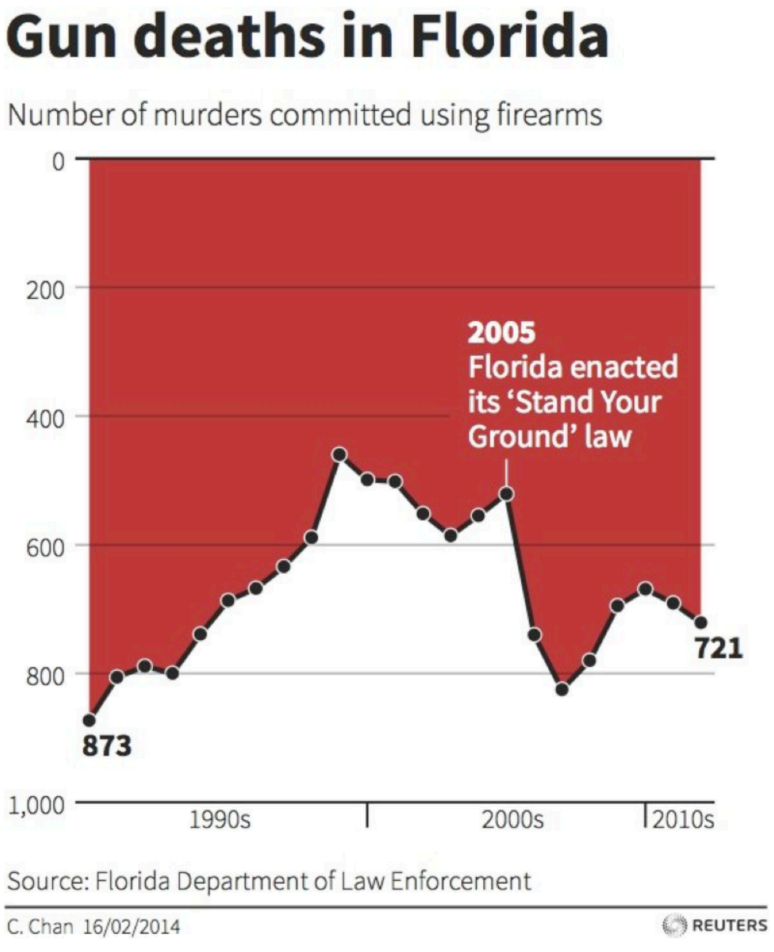
Better plot



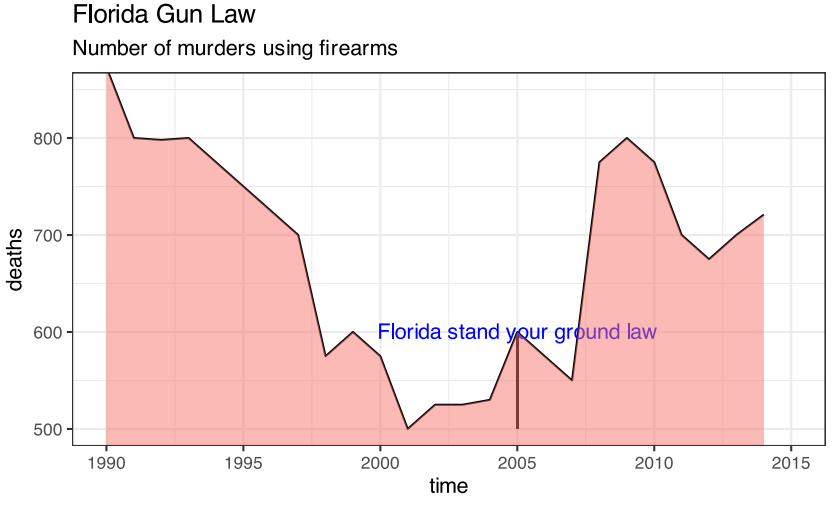
Barplots must include zero point to be properly interpreted

Deceptive graphs #3

Original plot



Better plot



Inverting an axis will obviously lead to confusion and misinterpretation

Deceptive graphs

Remember Hanlon's razor

Never attribute to malice that which can be adequately explained by stupidity

Accessible data visualizations

- We all process visualizations differently
- Do not rely on colour alone to convey a message
- The `viridis` package in R provides color palettes that are robust to colorblindness
- Digital Accessibility at McGill https://www.mcgill.ca/digital-accessibility-mcgill-university

Interested in learning more?

- Computational and Data Systems Intiative at McGill offers free workshop in R (and Python)

- Introductory and Intermediate Data Visualization in R
- Building Interactive Plots in R
- Creating a Shiny Dashboard in R

- CDSI Workshops