

Week 3: Graphics and Visualization

MATH-517 Statistical Computation and Visualization

Tomas Masak

Oct 07, 2022

Graphics

- presentation
 - result communication
 - decision making
- data insight
 - large data
 - detect patterns
 - find strange observations
- code debugging
 - input, output (even the code itself) is data

Graphics

“The simple graph has brought more information to the data analyst’s mind than any other device.” – John W. Tukey

“The greatest value of a picture is when it forces us to notice what we never expected to see.” – John W. Tukey

One can think of graphics (and also models, for that matter) as of a low-dimensional representation for data.

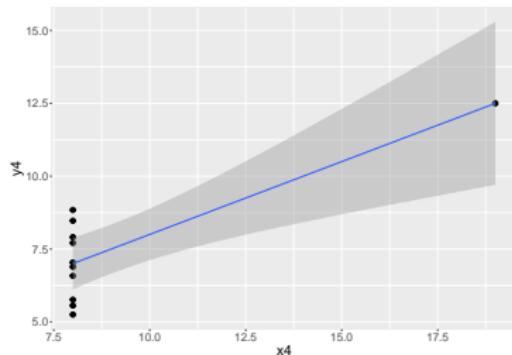
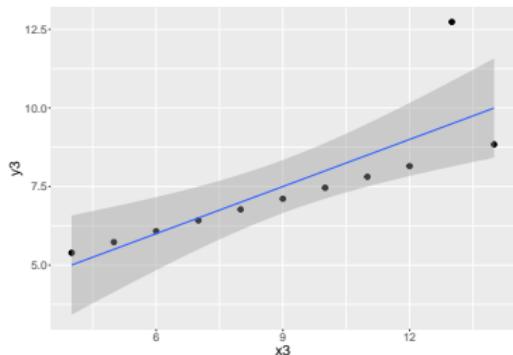
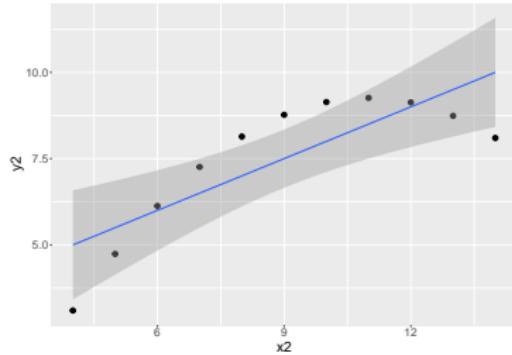
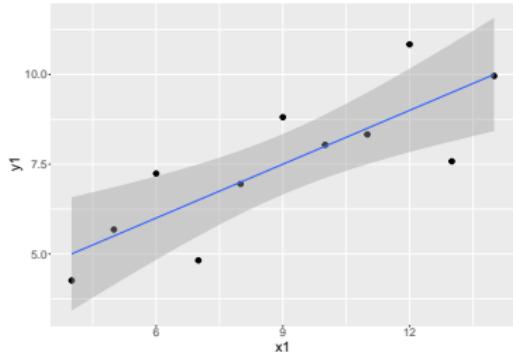
Anscombe's Quartet

Four data sets with

- one response variable y
- one regressor x

```
##          (Intercept)           x  R-squared
## lm1      3.000091  0.5000909 0.6665425
## lm2      3.000909  0.5000000 0.6662420
## lm3      3.002455  0.4997273 0.6663240
## lm4      3.001727  0.4999091 0.6667073
```

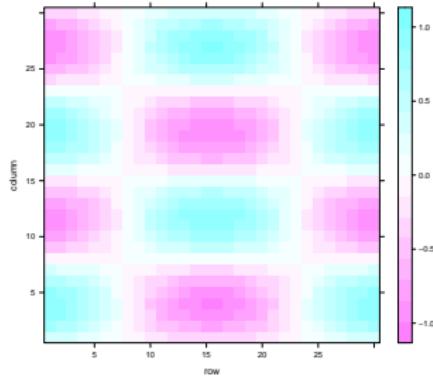
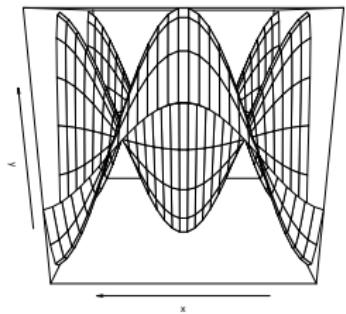
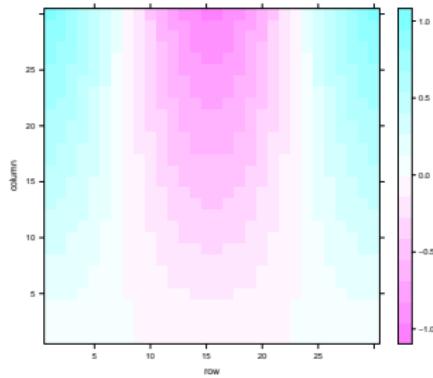
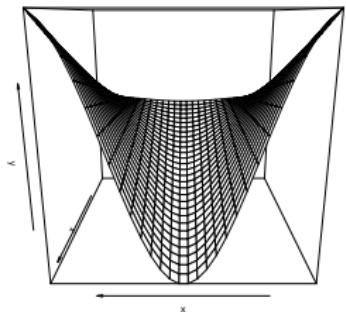
Anscombe's Quartet



Human Height



Plot Dimension



Scatterplot Extras

The following features (ggplot's arguments) for points (and similarly for lines)

- color
- shape
- size
- alpha (opacity/transparency)

can be used

- to include additional information (or dimensions, i.e. to include additional variables) in a scatterplot
- to combat overplotting
- or simply to make the plot nicer (i.e. used subjectively)

Available Shapes

pch = _

1 ○ 6 ▽ 11 ✕ 16 ● 21 ○

2 △ 7 ✷ 12 ✸ 17 ▲ 22 □

3 + 8 * 13 ✷ 18 ◆ 23 ◇

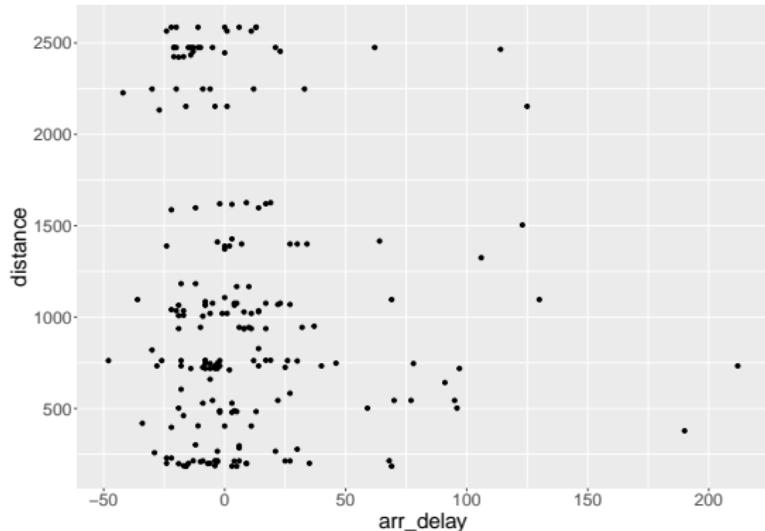
4 ✖ 9 ✦ 14 ✸ 19 ● 24 △

5 ◇ 10 ⊕ 15 ■ 20 • 25 ▽

- pch is the base R argument, for ggplot one instead passes values into `scale_shape_manual()`
- all shapes have attribute color, only some fill

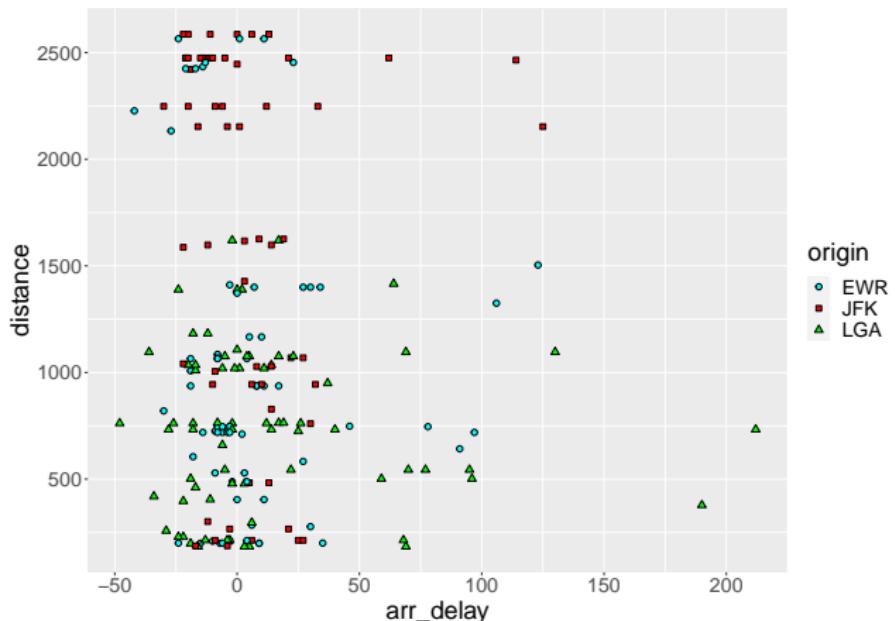
Scatterplot: plain

```
library(data.table)
set.seed(517)
flights <- fread("https://raw.githubusercontent.com/Rdatatable/data.table/m
  slice_sample(n=200)
ggplot(data = flights,
       mapping = aes(x = arr_delay, y = distance)) +
  geom_point()
```



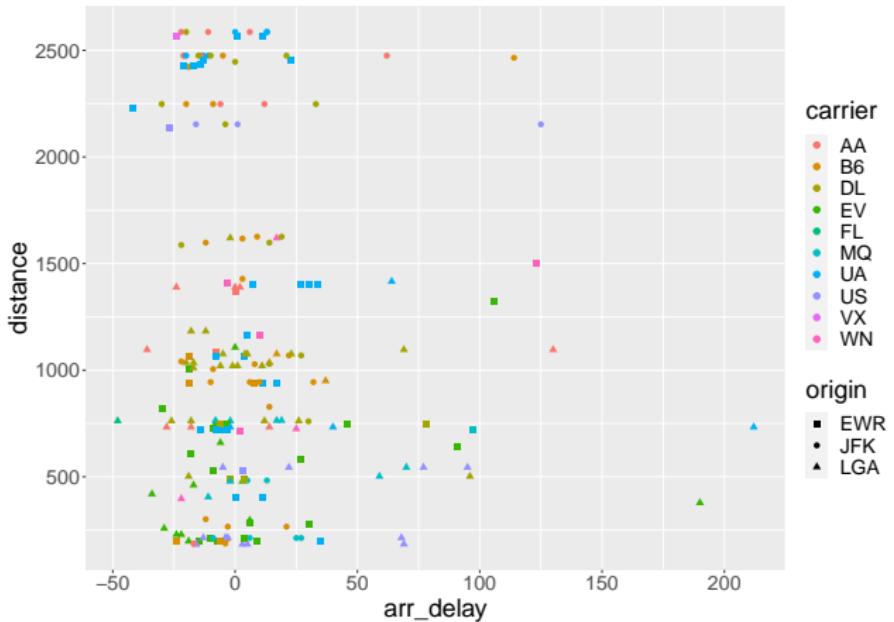
Scatterplot: shape + color

```
ggplot(data = flights, mapping = aes(x = arr_delay, y = distance,  
                                      shape = origin, fill=origin)) +  
  geom_point(size=2) +  
  scale_fill_manual(values = c("cyan","red","green")) +  
  scale_shape_manual(values = c(21,22,24))
```



Scatterplot: shape × color

```
ggplot(data = flights, mapping = aes(x = arr_delay, y = distance,  
                                      shape = origin, color=carrier)) +  
  geom_point(size=2) + scale_shape_manual(values = c(15,16,17))
```



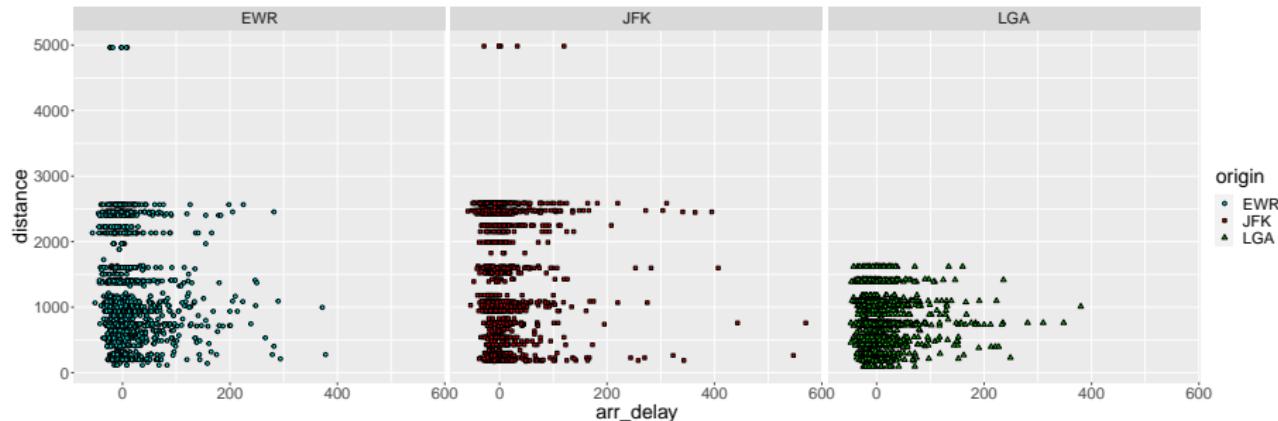
Overplotting

```
set.seed(517)
flights <- fread("https://raw.githubusercontent.com/Rdatatable/data.table/m
  slice_sample(n=5000)
ggplot(data = flights, mapping = aes(x = arr_delay, y = distance,
                                      shape = origin, fill=origin)) +
  geom_point(size=2) +
  scale_fill_manual(values = c("cyan","red","green")) +
  scale_shape_manual(values = c(21,22,24))
```



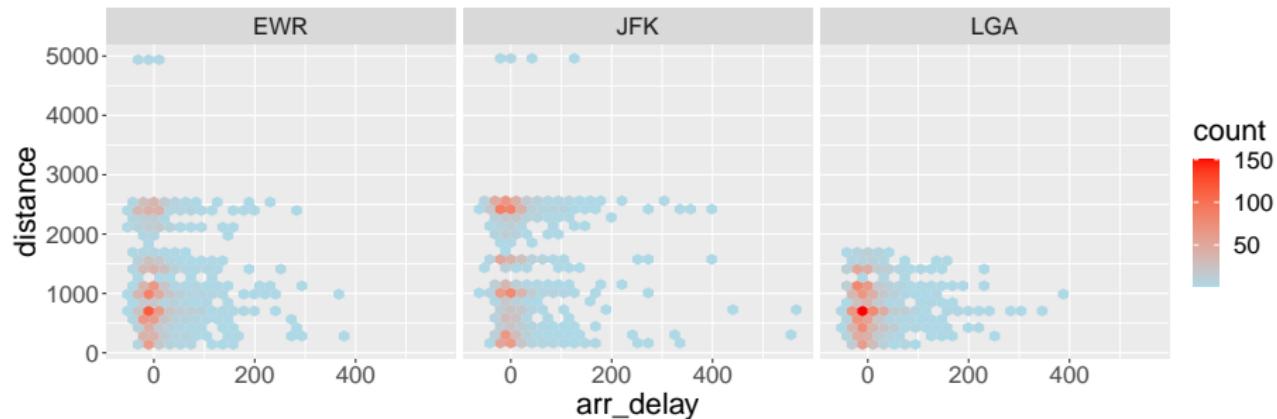
Overplotting Remedy 1/2

```
ggplot(data = flights, mapping = aes(x = arr_delay, y = distance,  
                                      shape = origin, fill=origin)) +  
  geom_point() +  
  scale_fill_manual(values = c("cyan","red","green")) +  
  scale_shape_manual(values = c(21,22,24)) +  
  facet_wrap(~origin)
```



Overplotting Remedy

```
ggplot(data = flights, mapping = aes(x = arr_delay, y = distance)) +  
  stat_binhex() +  
  scale_fill_gradient(low = "lightblue", high = "red") +  
  facet_wrap(~origin)
```



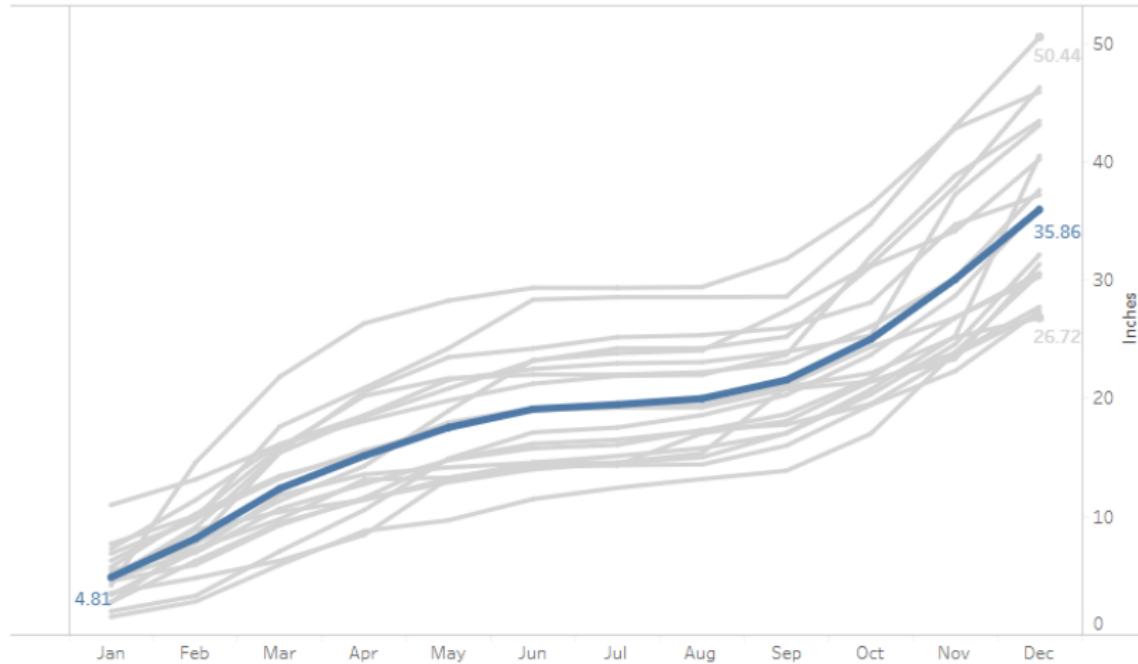
Overplotting Remedies

- reducing point size
- split a single plot into multiple (done above)
- jittering
- opacity/transparency
- binning (done above via `stat_binhex()`)
- tiles
- subsampling (done above above)

Many of these options clash with clarity and space requirements or with each other (e.g. transparency distorts colors).

Visualizing Variance

Average annual rainfall in Portland

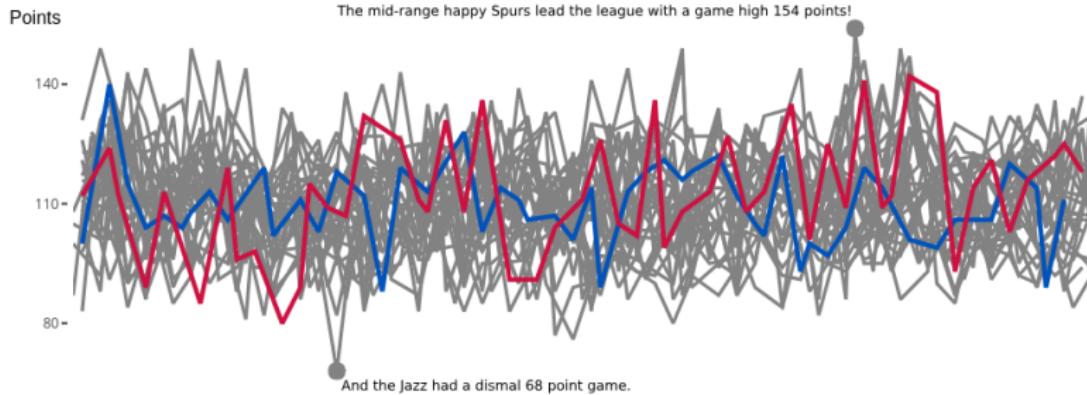


Portland (Oregon) saw an average rainfall of 35.86 inches a year between 2010 and 2018. The wettest year (2012, 50.44") was followed by the driest year (2013, 26.72") during this time frame.

Source: National Weather Service Forecast Office <https://w2.weather.gov/climate/xmacis.php?wfo=pgr>

Visualizing Variance (Bad)

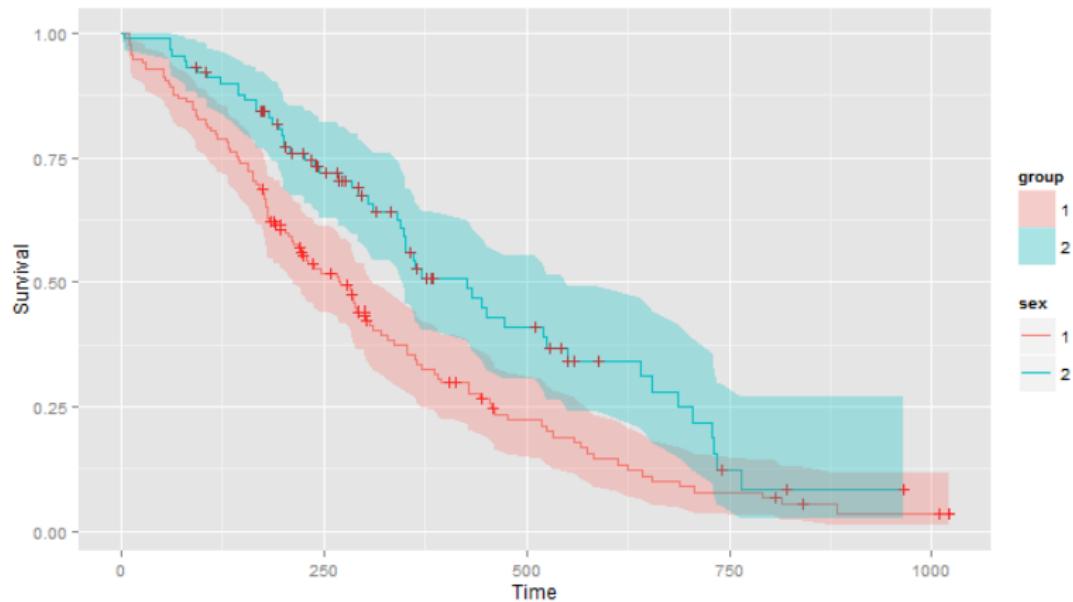
The **Houston Rockets** have the highest game-to-game point variance and the **Dallas Mavericks** have been the most consistent in the 2018-19 NBA season



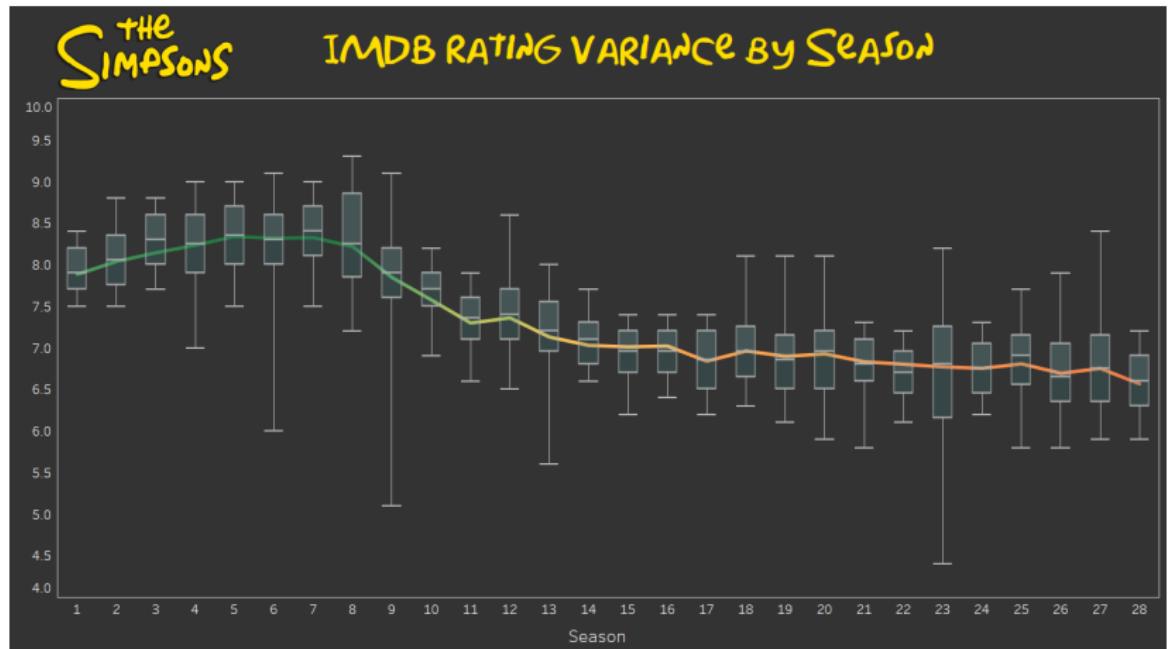
Source: stats.nba.com
API: https://github.com/swar/nba_api
Edited with: <https://www.photopea.com/>

- the grey lines should be much thinner and colored so we can relate them to the colored means (and x-axis added)

Visualizing Variance



Visualizing Variance



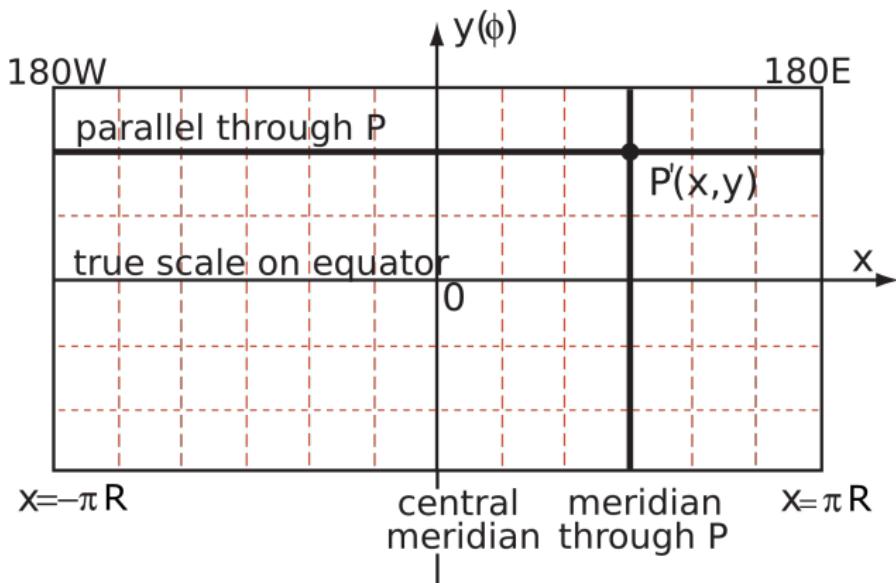
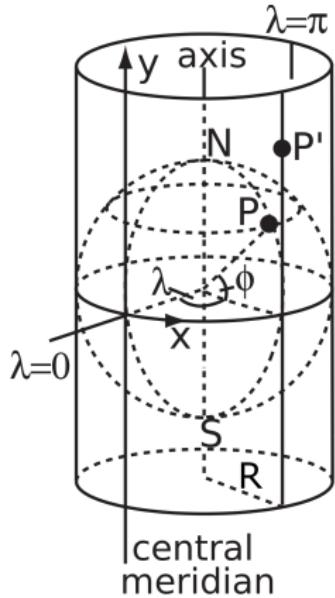
Spatial Data

Complicated due to different

- mathematical representations
- data structures
- data sources
- data processing packages
 - projections (sphere? plane?)
- visualization packages

(Not so) short course about visualizing spatial data [here](#) (only if interested).

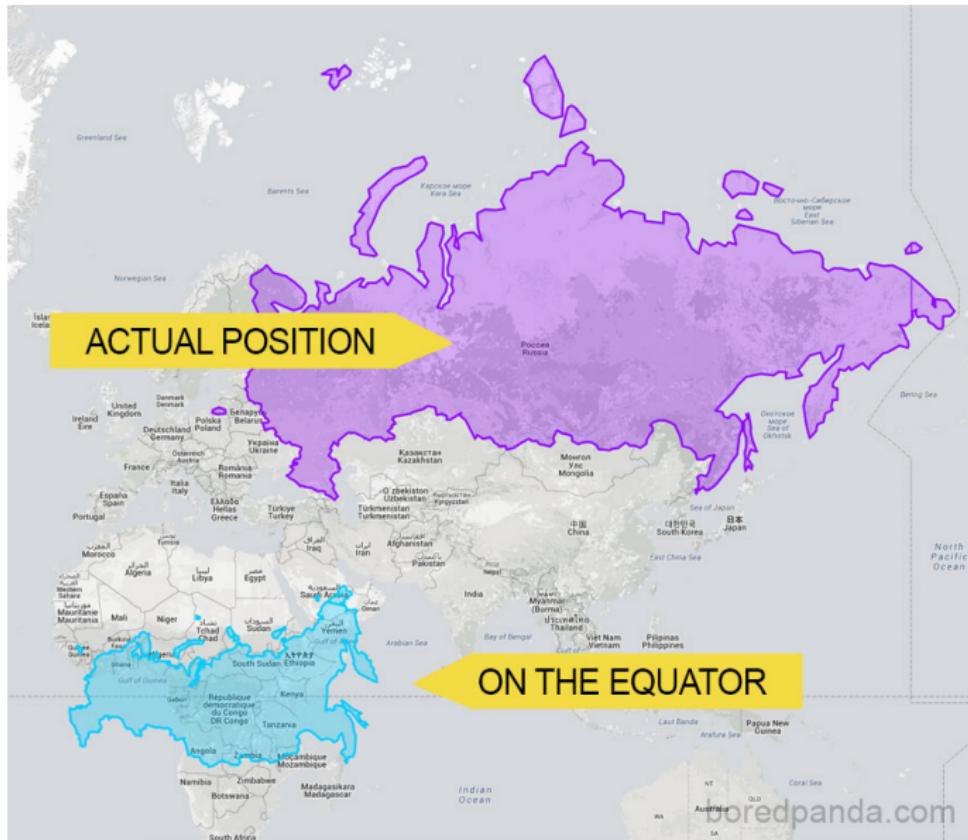
Mercator Projection



source: wiki

Two sources of distortion (earth is not a sphere and the projection itself), the second one visualized [here](#).

Russia's True Size



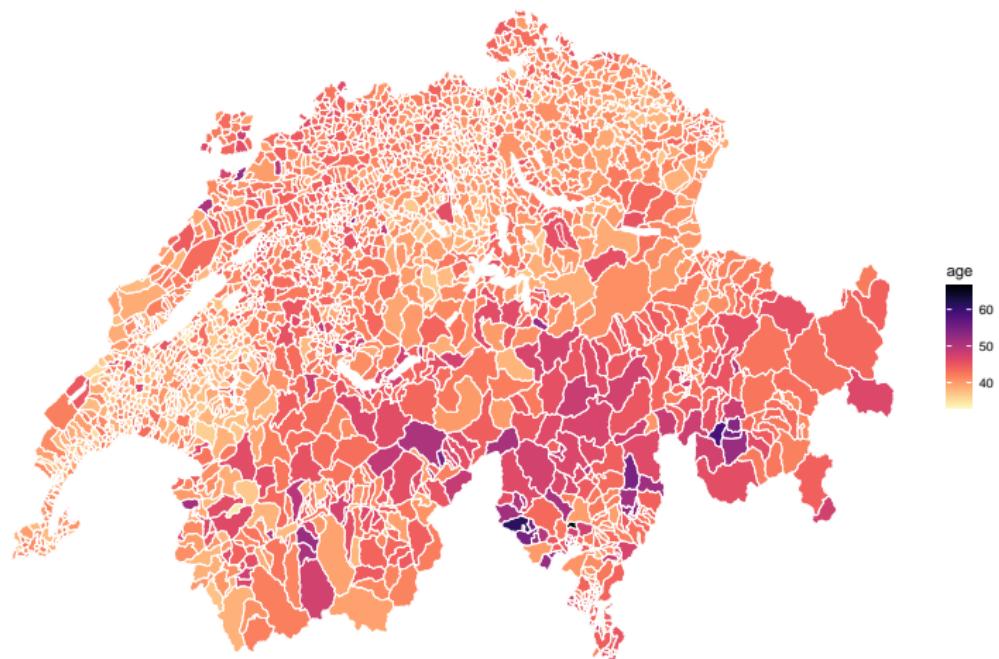
Geospatial Heatmap

```
library(ggswissmaps)
counties <- shp_df[[4]] # resolution on the level of ZIP codes

avg_age <- read.csv("https://raw.githubusercontent.com/grssnbchr/thematic-maps/master/data/age.csv")
names(avg_age)[1] <- "id"
names(avg_age)[3] <- "age"
counties <- counties %>% mutate(id = as.numeric(id))
newdat <- inner_join(counties, avg_age, by="id")
library(viridis)
ggplot(data=newdat) +
  geom_polygon(aes(x=long, y=lat, group=group, fill=age)) +
  geom_path(aes(x = long, y = lat, group = group),
            color = "white", size = 0.1) +
  scale_fill_viridis(option = "magma", direction = -1) +
  theme_void()
```

- here `counties` is a data frame, longitude and latitude specify border points and Swiss counties are polygons (convex hulls of the border points)
- `group` specifies which border points belong to which county
- `try plot(counties$long, counties$lat, type="l")`

Geospatial Heatmap



inspired by [this blogpost](#)

Good Visualization Practices

- provide context (in text **and** in caption)
- seek simplicity, clarity, etc.
- gray scale often preferable
 - color-blindness (friendly palettes, e.g. [Cools](#))
- axes (scale, gaps, etc.)
 - text of appropriate size
- publication-specific conditions
- be artistic!
 - sometimes bend the rules (responsibly and justifiably)

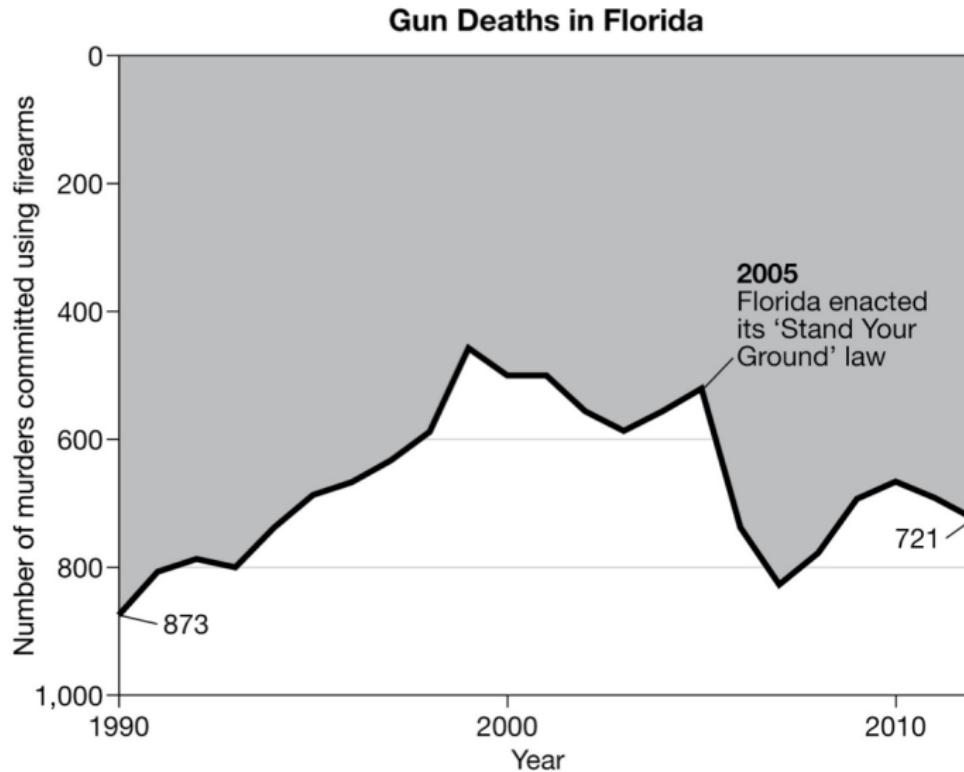
Find inspiration in [The R Graph Gallery](#).

Beware when exporting graphics.

Section 1

Bad Visualization Practices

Reverted Axis

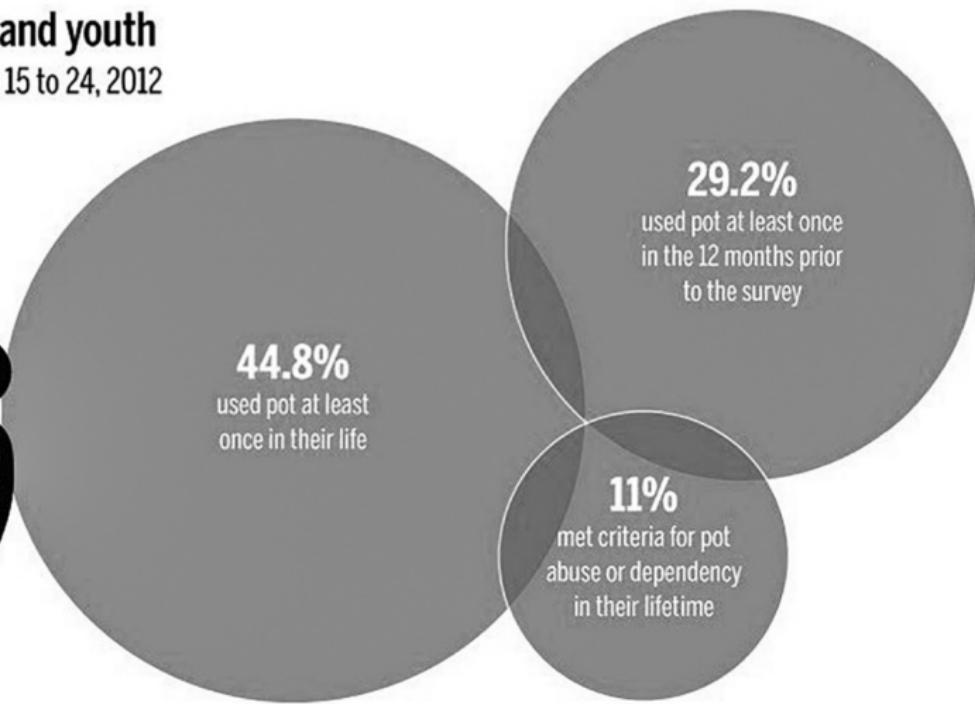


Source: Florida Department of Law Enforcement

False Venn's Diagrams

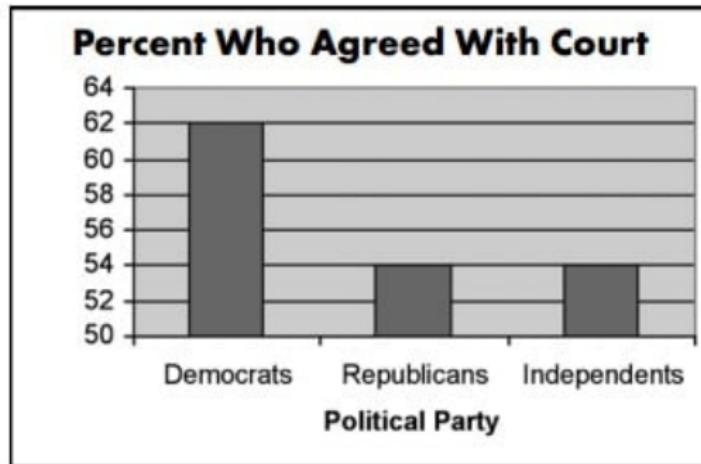
Marijuana and youth

Canadians age 15 to 24, 2012



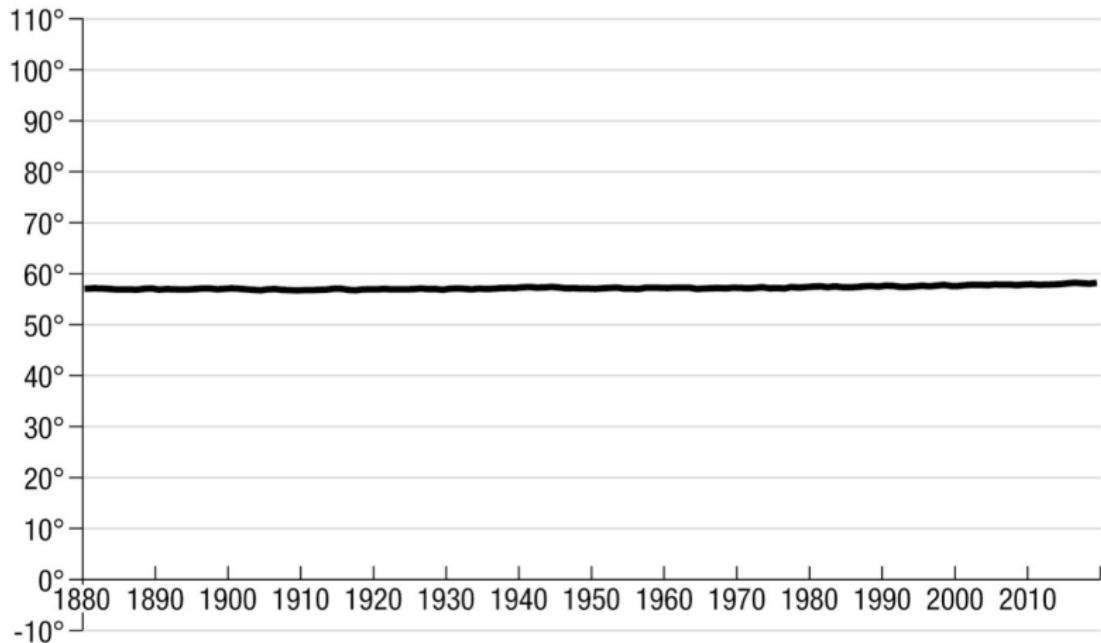
SOURCE: STATISTICS CANADA

Missing Baseline



Axis Starting at Zero

Average Annual Global Temperature in Fahrenheit, 1880–2019



Average Global Temperature by Year



Combined Effects

Women's earnings as a percentage of white men's earnings



hillaryclinton • Follow

...



ayolucasss @eimear.ml you have no idea how they've reached this conclusion. Do you even know how they go about calculating this? Of course not.

241w



ayolucasss @eimear.ml THIS IS ON HILLARY'S INSTAGRAM 😂😂 And that's all your credibility out the window.

241w



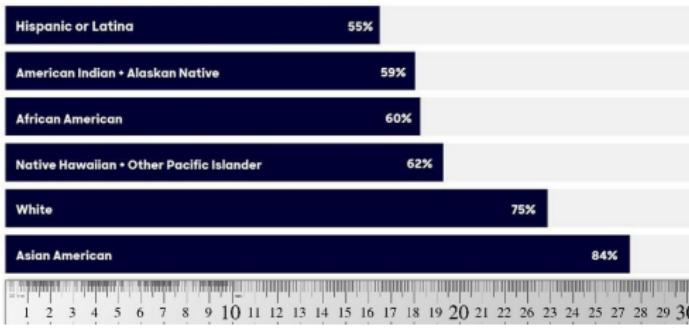
17,708 likes

APRIL 12, 2016

Comments on this post have been limited.

... plus just Cheating

Women's earnings as a percentage of white men's earnings



hillaryclinton • Follow

...



ayolucasss @eimear.ml you have no idea how they've reached this conclusion. Do you even know how they go about calculating this? Of course not.

241w



ayolucasss @eimear.ml THIS IS ON HILLARY'S INSTAGRAM 😂😂 And that's all your credibility out the window.

241w



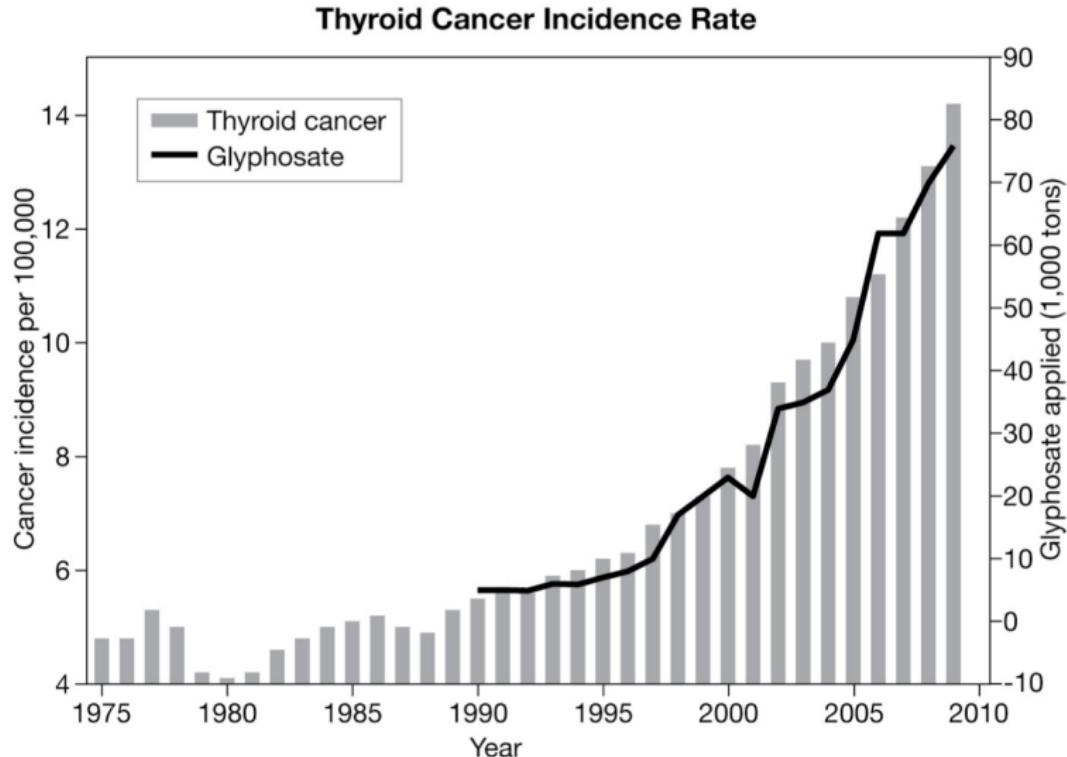
17,708 likes

APRIL 12, 2016



Comments on this post have been limited.

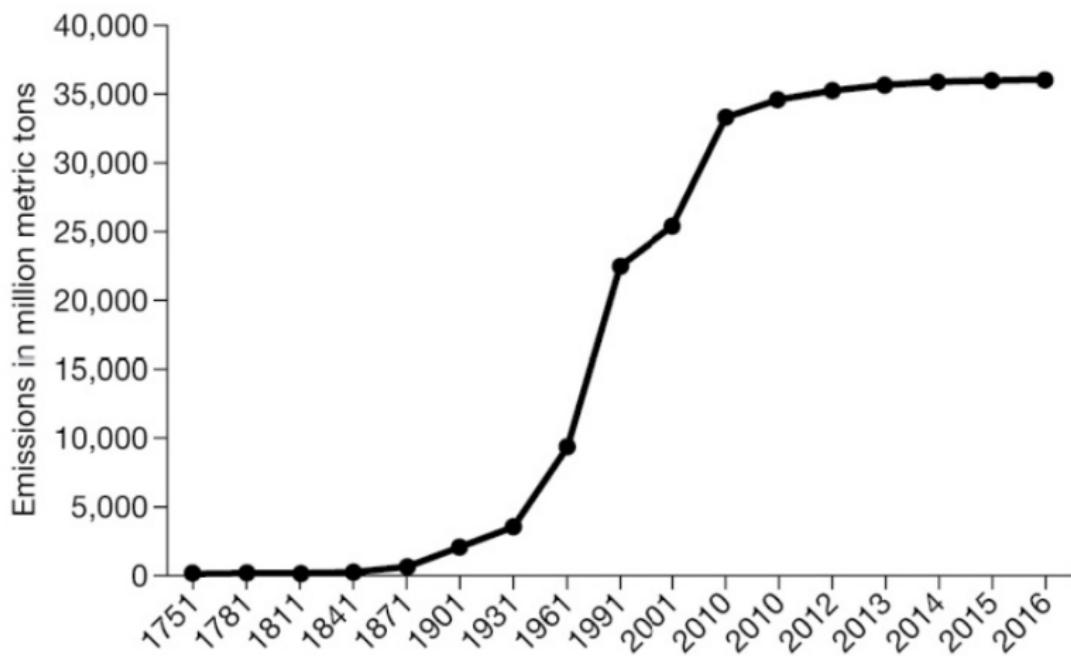
Double Axes



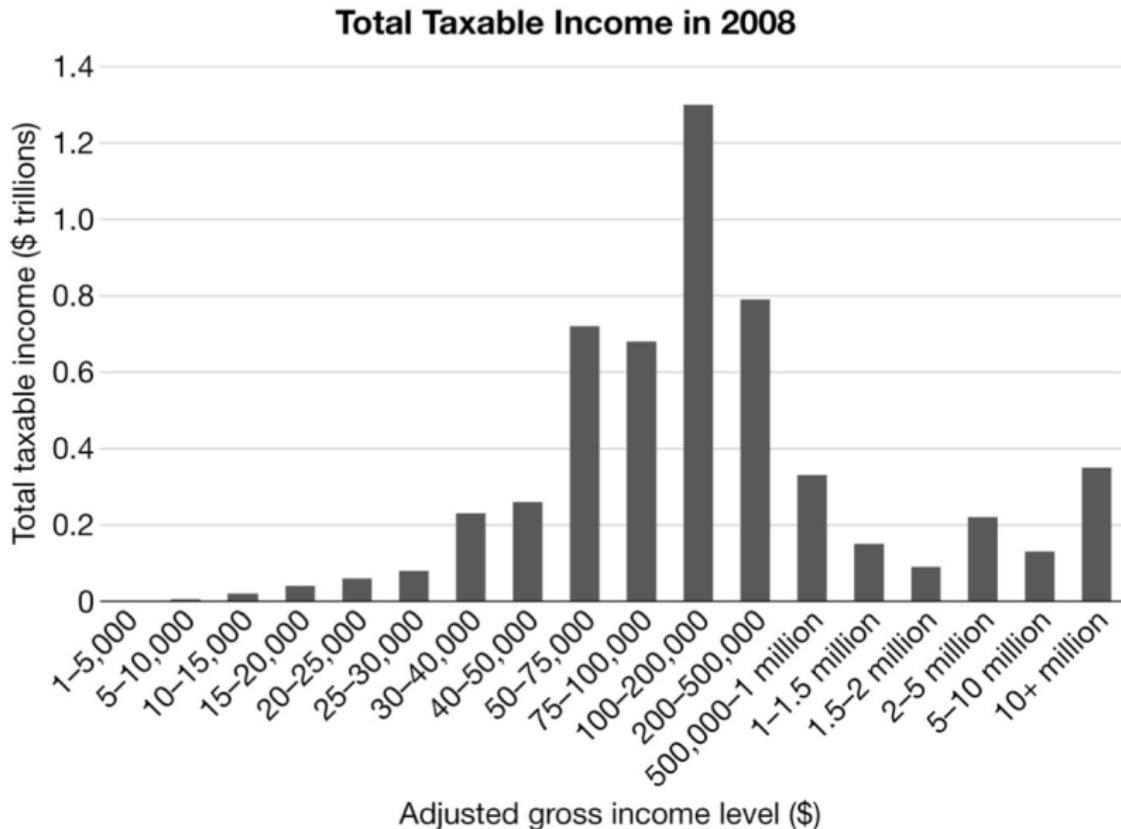
- This is actually quite good, but double axes are usually problematic.

Tweaking Axis

Carbon Dioxide Emissions from Global Fossil Fuel Combustion and Industrial Processes, 1751–2016

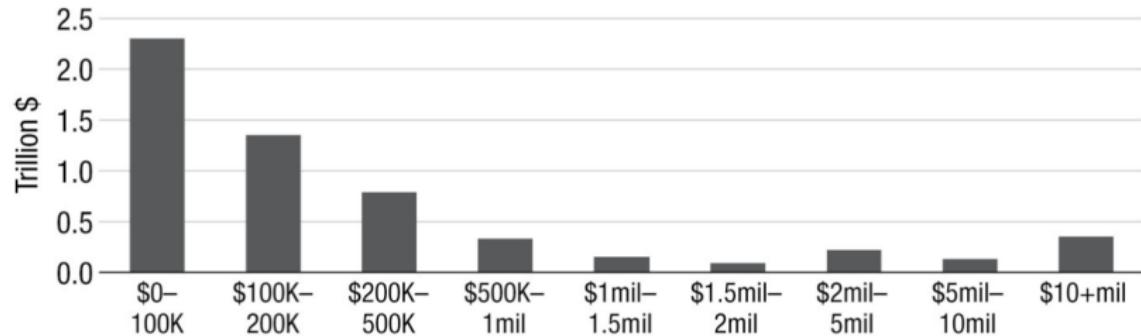


Binning

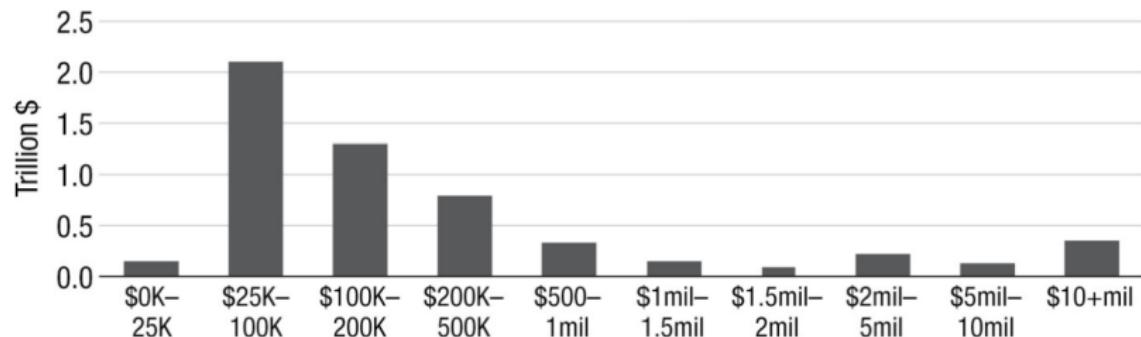


Different Kinds of Binning

Tax the Poor!

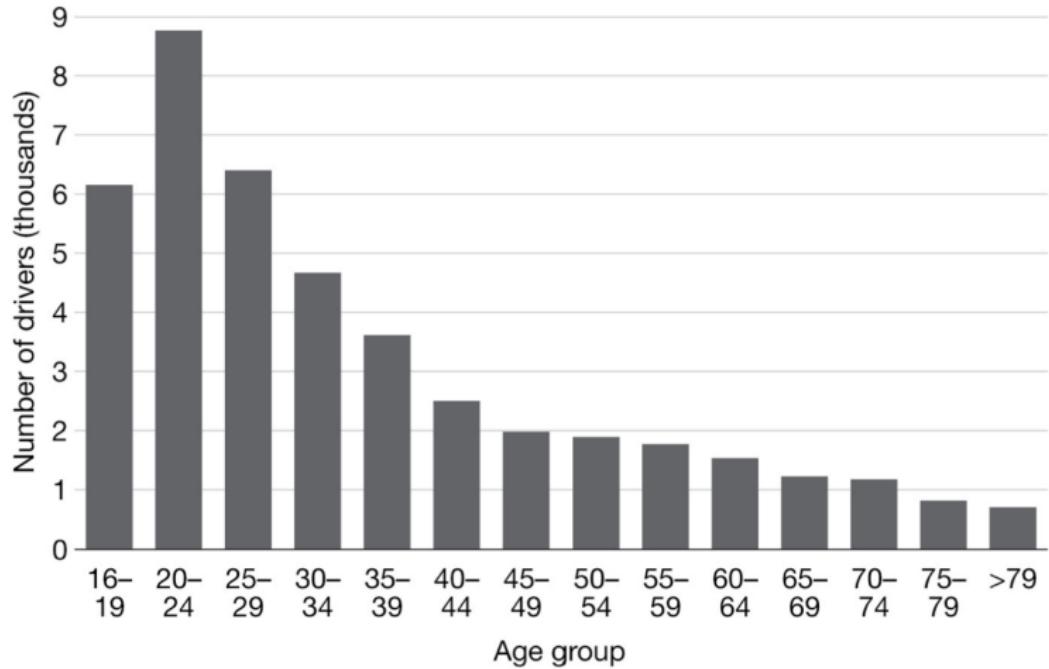


Tax the Middle Class!

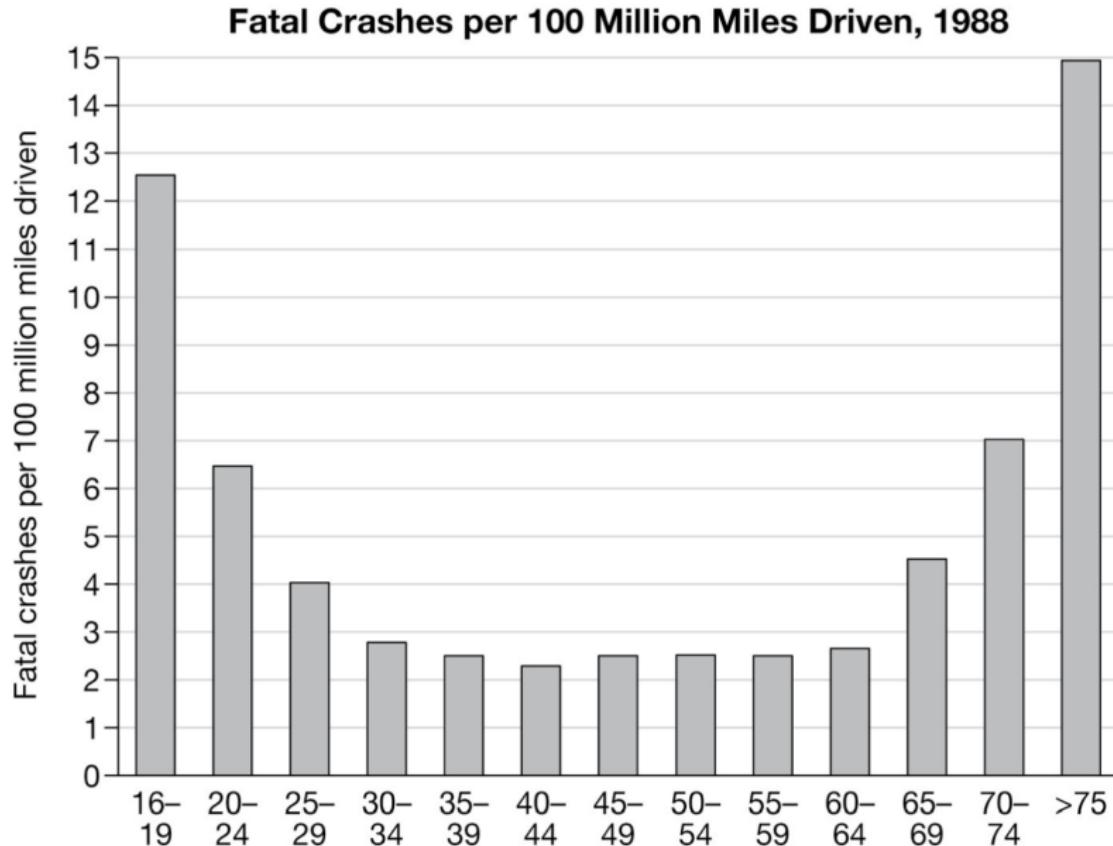


Total

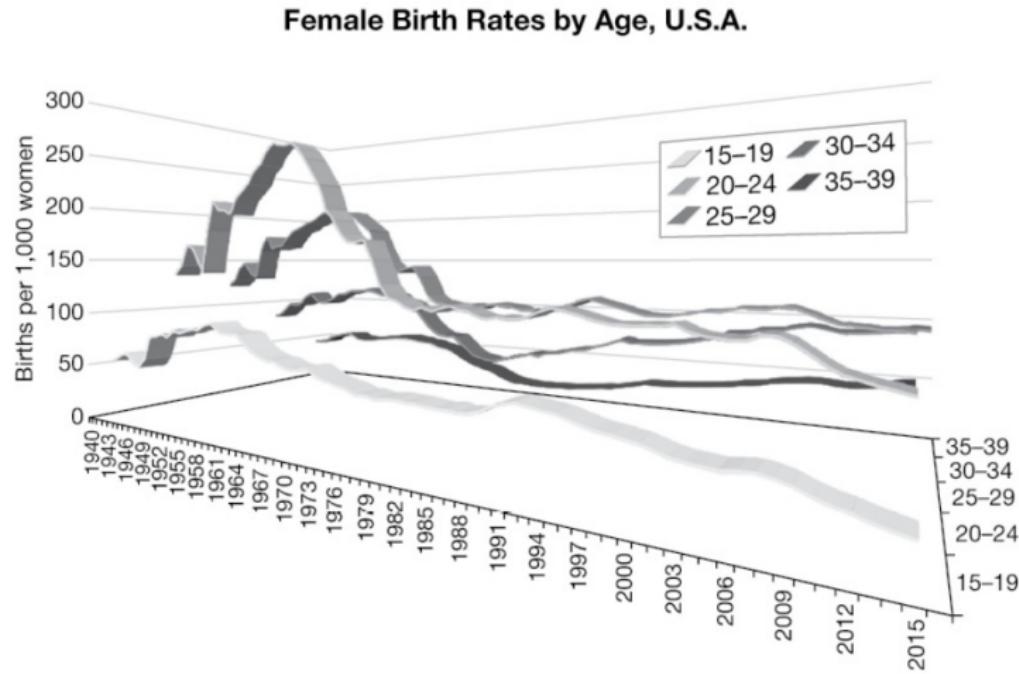
Number of Drivers in Fatal Crashes, 1988



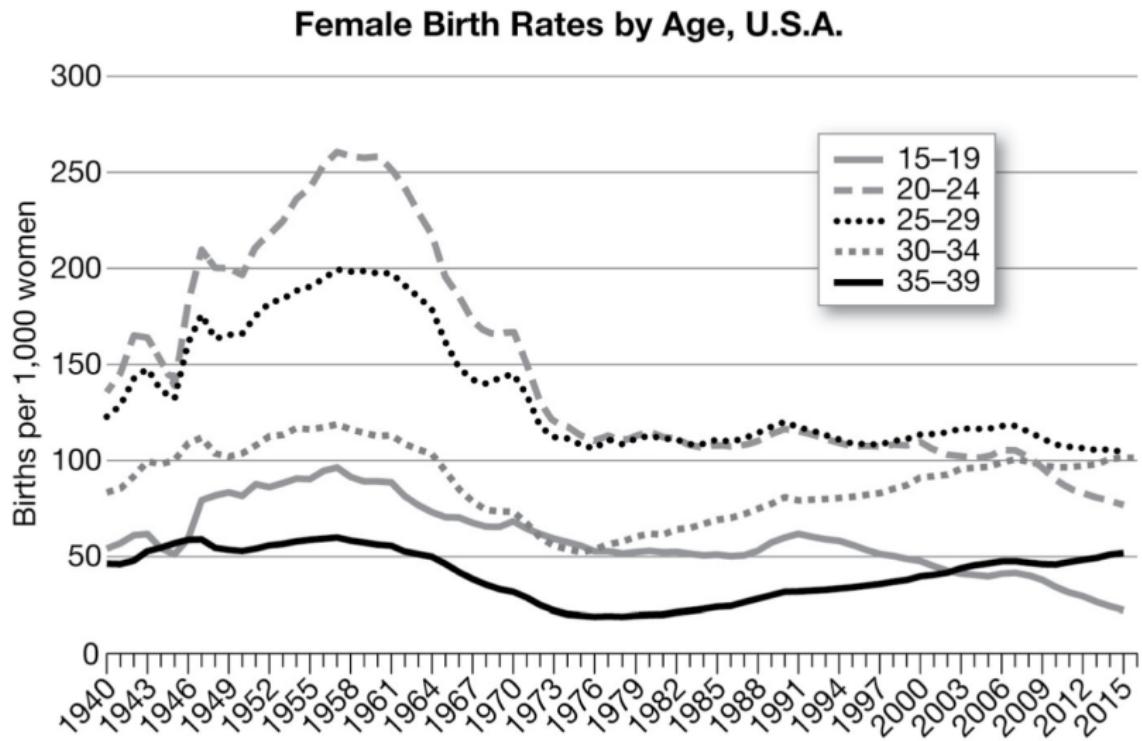
Relative



Useless 3D

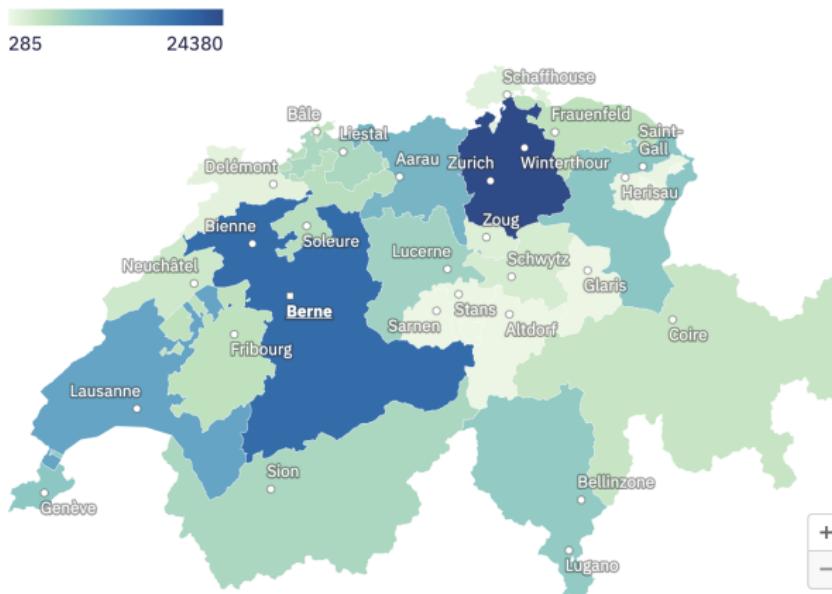


Better 2D



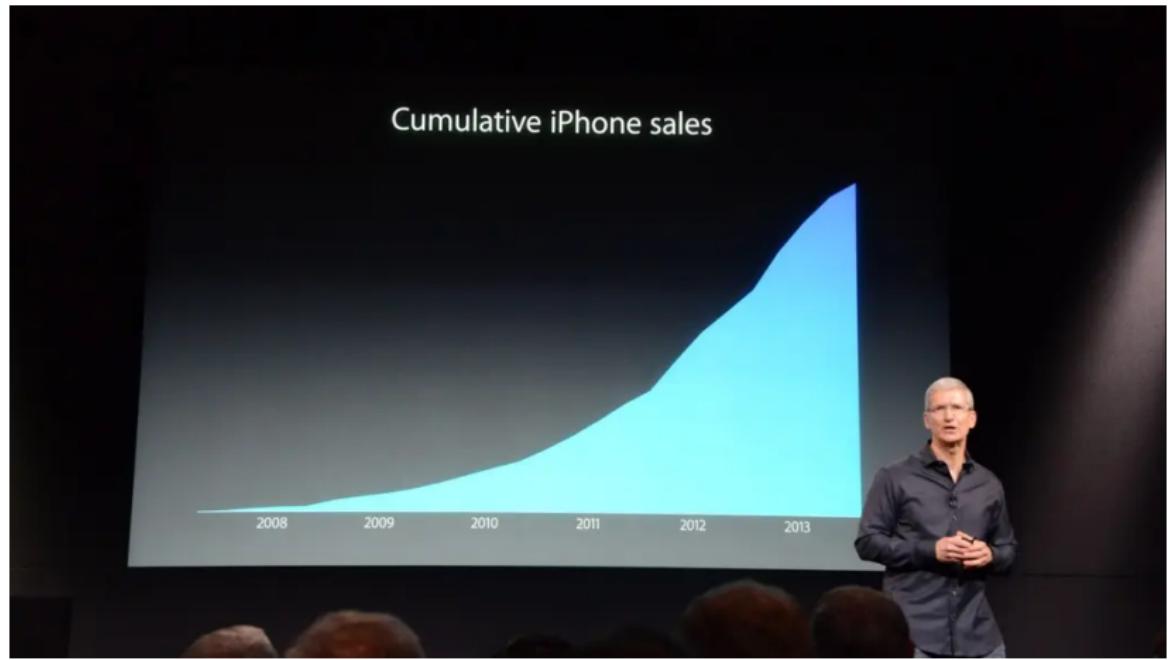
Total vs. Relative Again

Nombre de personnes atteintes de démence dans les différents cantons suisses



Carte: G. Laplace. Nombre exact par canton en se positionnant dessus avec la souris; Source: Faits et chiffres (Alzheimer Suisse, 2021); [Récupérer les données](#)

Missing Axis & Misguidance



One more Scatterplot Example

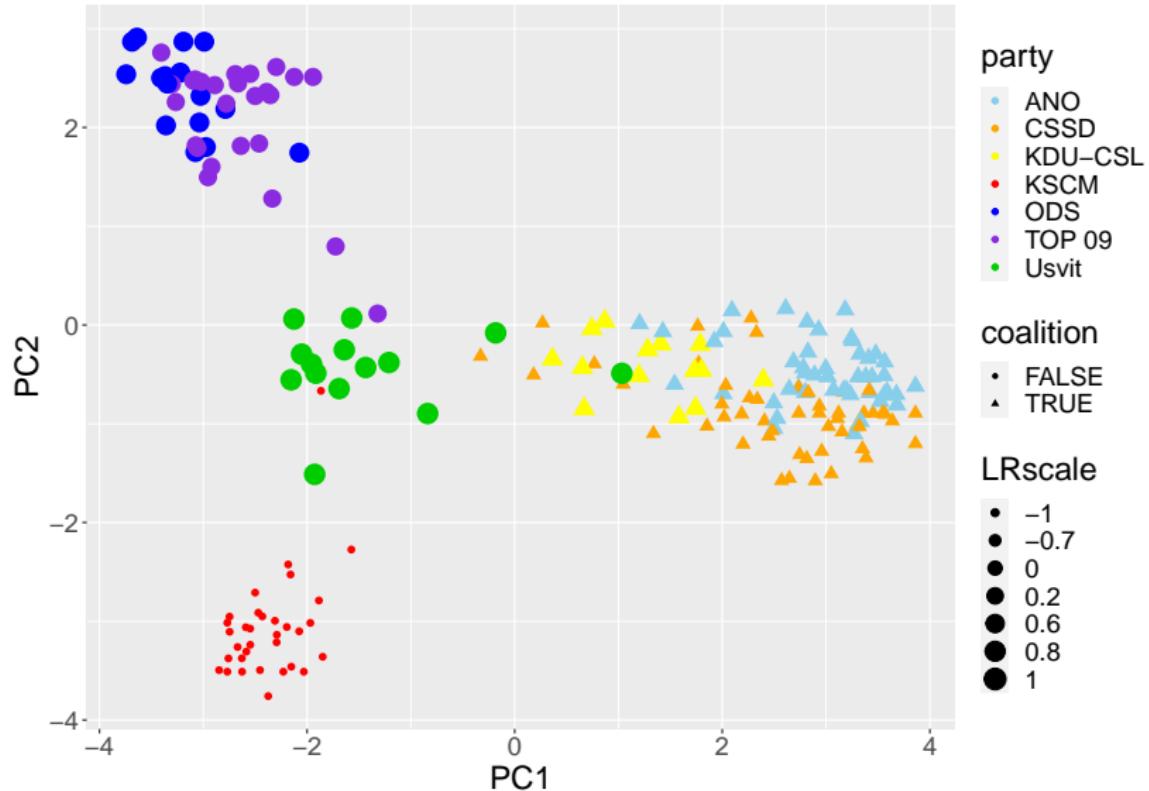
- voting records of the members of the Czech parliament in 2015
 - $N = 200$ members of the parliament
 - certain number of votings during that period ($x_{nj} \in \{-1, 0, 1\}$ depending on whether n -th member voted for the j -th voting no/abstain/yes), but **PCA** applied and only first 2 components kept
- additional information about party affiliation of the members
 - party affiliation (7 parties in total)
 - government coalition affiliation (T/F)
 - left-wing/right-wing scale LRscale (on the party level)

```
## 'data.frame':    200 obs. of  6 variables:  
##   $ PC1        : num  -1.93 -2.16 -3.04 2.74 -1.94 ...  
##   $ PC2        : num  -1.512 -2.527 2.051 -0.892 2.512 ...  
##   $ party      : chr  "Usvit" "KSCM" "ODS" "CSSD" ...  
##   $ party_color: chr  "green3" "red" "blue" "orange" ...  
##   $ coalition  : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 2  
##   $ LRscale     : Factor w/ 7 levels "-1","-0.7","0",...: 7 1
```

One more Scatterplot Example

```
ggplot(data = cz, mapping = aes(x = PC1, y = PC2, color=party,  
                                 shape=coalition, size=Lscale)) +  
  geom_point() +  
  scale_color_manual(values=c("skyblue","orange","yellow",  
                            "red","blue","blueviolet","green3")) +  
  scale_shape_manual(values=c(16,17))
```

One more Scatterplot Example



Assignment

Small project [20%]. Deadline on Week 6.

The goal of this project is *data exploration*. Find an interesting (in the sense it interests you!) data set and

- explore the data
- describe the data
- visualize the data
- lay out some questions about the data based on your exploration
- use more detailed visualization techniques to hint answers

The first step should be done individually. Then you can form groups of 2-3 and pick up the most interesting data set and do the rest. See [Course organization](#) for more details.

Exercise

- ① Some data repositories listed on following slides will be explored.
- ② Consider the Simpsons IMDB ratings plot above. Choose say 3 of your favorite TV shows, download data from the [IMDB database](#) and produce a plot similar to the Simpsons plot above (i.e. variances captured with boxplots) using `ggplot()` depicting all 3 TV shows.
- ③ Find and obtain spatial boundary files of administrative regions of your home country (or use some other country, if problematic). Produce a geospatial heatmap of the country, using some external variable for the color fill argument (e.g. specific election results or election participation, mortality rates, etc.).
- ④ Choose a scatterplot above (say the one on slide 10) and try to reproduce it using base R functions `plot()` instead of `ggplot()` and add legend manually using `legend()`.

Exercise: Some Links to Open Data

[fivethirtyeight](#): article data of Nate Silver's data journalism platform freely available (see also R package [fivethirtyeight](#))

[data-is-plural](#): weekly newsletter of datasets by Jeremy Singer-Vine

[re3data](#): Registry of research data repositories

[openml datasets](#): many uniformly formatted datasets for training machine learning models – however, not always good descriptions available

[Worldbank Datacatalog](#): the World Bank data catalogue

[UK Data Service](#): UK's largest collection of social, economic and population data resources (filter for open data) or also [data.gov.uk](#)

[ICPSR](#): unit within the Institute for Social Research at the University of Michigan, social and behavioral research. In particular including [replication datasets](#) for published studies.

[govdata](#): Open Government - German administrative data freely accessible

Exercise: Some more

[gapminder](#): “an independent educational non-profit fighting global misconceptions”; collection and visualization of datasets concerning global development

[nature.com](#): peer-reviewed, open-access journal for descriptions of datasets (broad range of natural science disciplines)

[NIH \(National Institute of Health\) Data Sharing Repositories](#): overview on different thematically sorted medical databases

[UCI Machine Learning Repository](#) or the new [beta version](#): containing various datasets – however, sometimes with a little few description

[data.bris Research Data Repository](#): Data repository of the University of Bristol

[bellingcat TikTok Hashtag analysis tool](#): Didn't try, don't know how easy.
... *no systematic selection. Much more out there*