# Calling Bullsh*t on Data Viz

CM4125 – Week 4

# Announcements

- Datacamp Academic Group

- (More) Coursework Clarification
  - Word count limits
  - Video from last week is embedded in Moodle
  - I will explain the clustering plot at the end of this lecture
  - I will attend more questions after the lab
    - Marking grid (i.e. what are you expected to do)

# What this lecture is about?

- Based on Week 6 of the Calling Bullsh*t course of the University of Washington by Carl T. Bergstrom and Jevin West
  - https://www.callingbullshit.org/syllabus.html#Visual

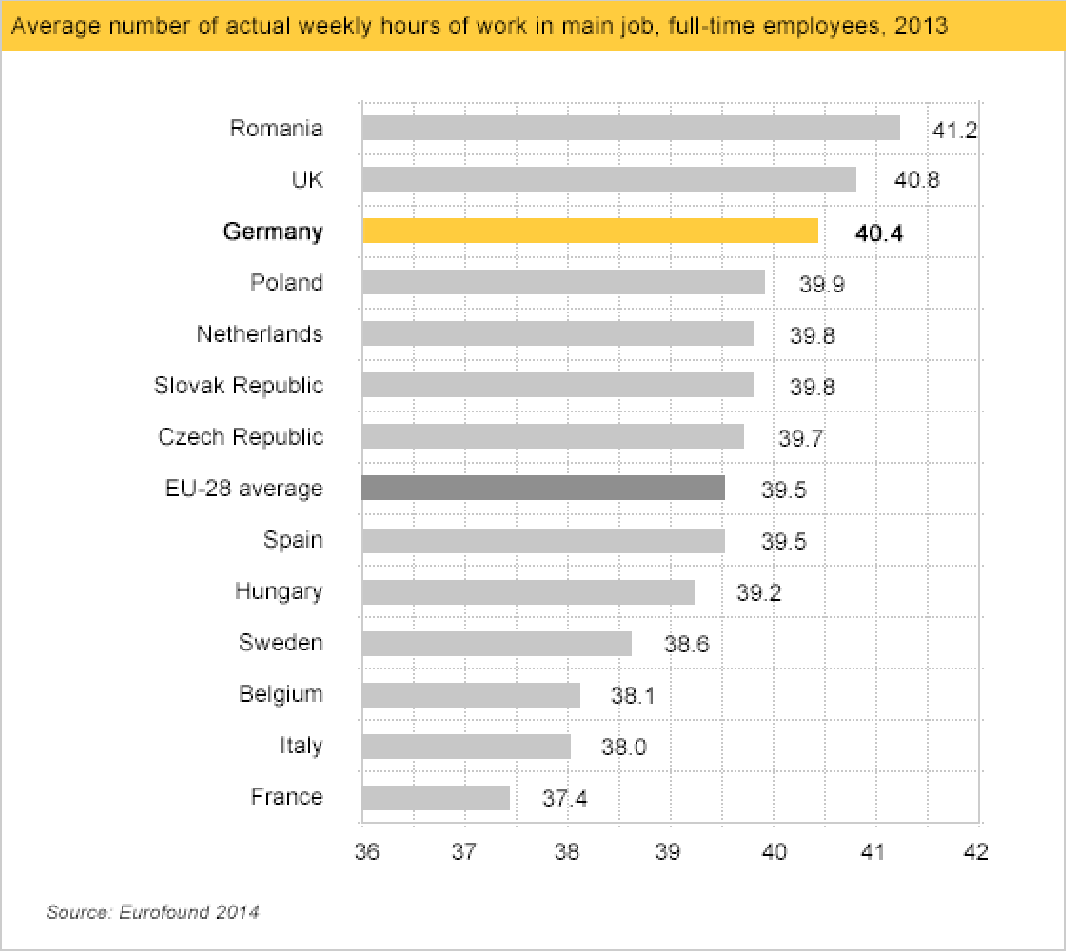
- Supplementary reading:
  - Alberto Cairo (2019) *How Charts Lie: Getting Smarter about Visual Information*. W.W. Norton and Company.
  - Edward Tufte (1983) *The Visual Display of Quantitative Information*. Chapters 2 (Graphical integrity) and 5 (Chartjunk: vibrations, grids, and ducks).

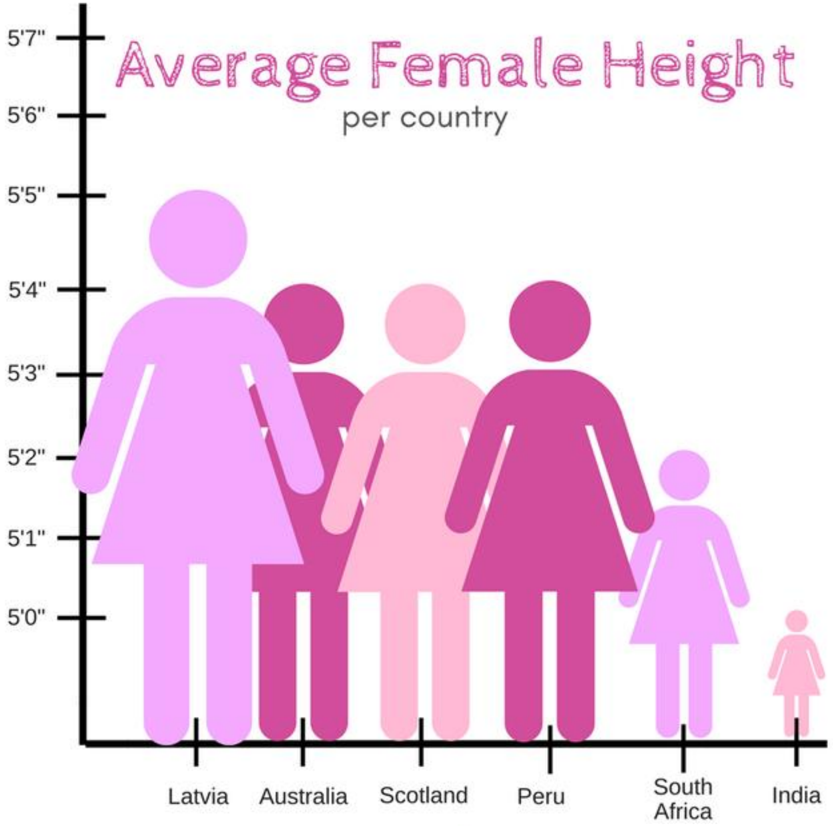
# Misleading axis

Charts can conceal or mislead if axes are set incorrectly

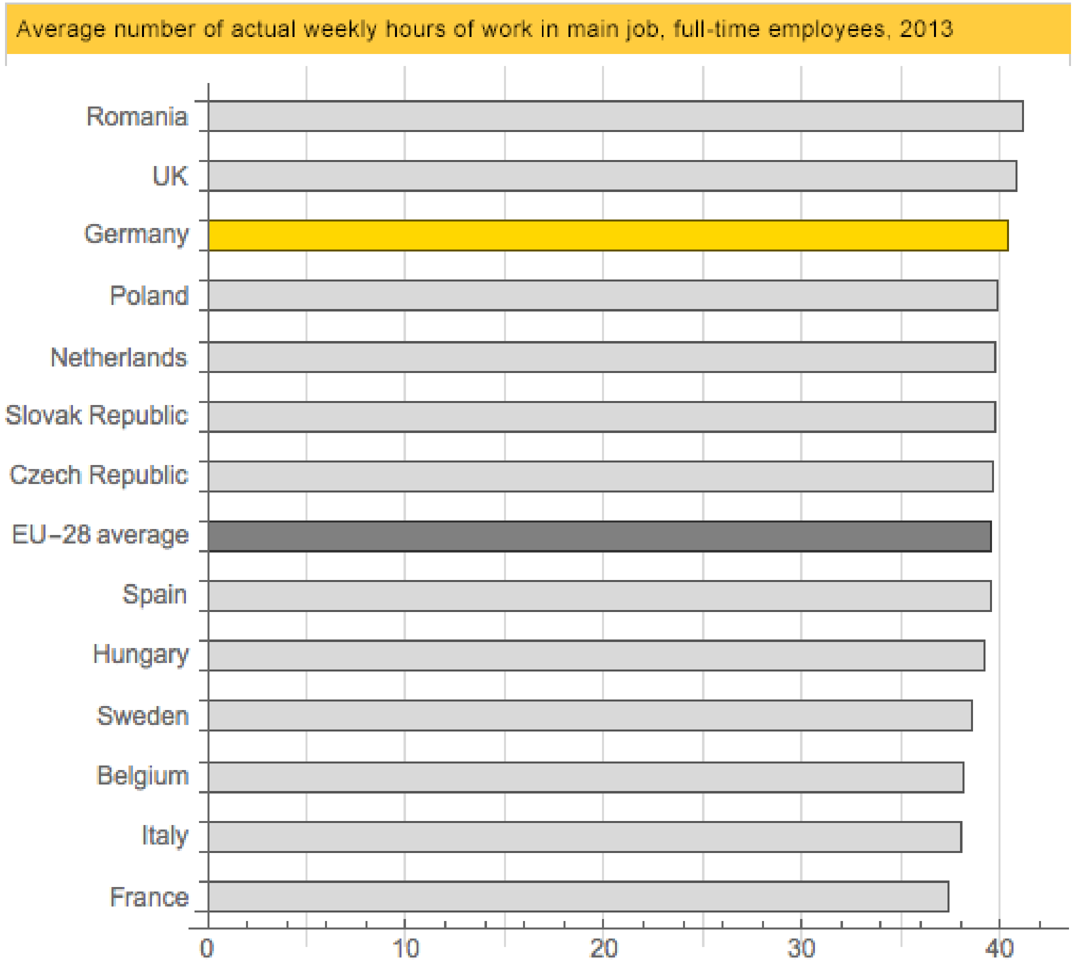
How to choose the range and scale of axes in a graph?

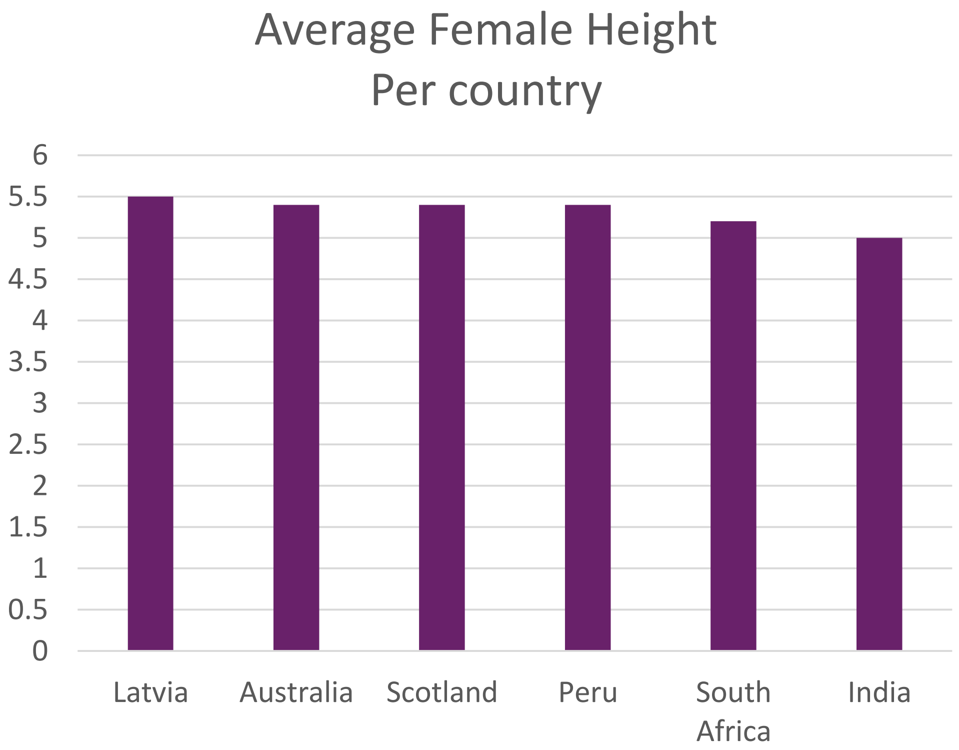
# 1. Bar chart axes **should** include zero



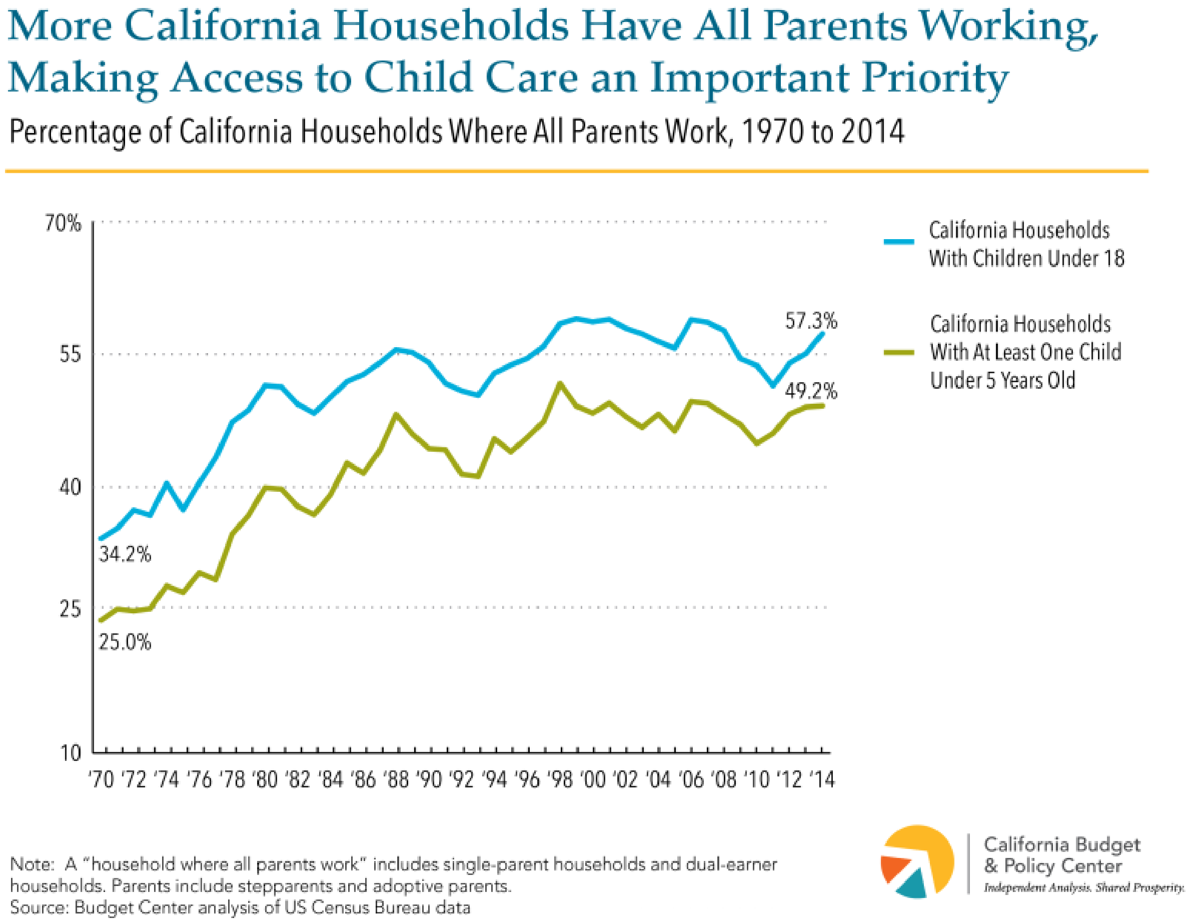


# 1. Bar chart axes **should** include zero





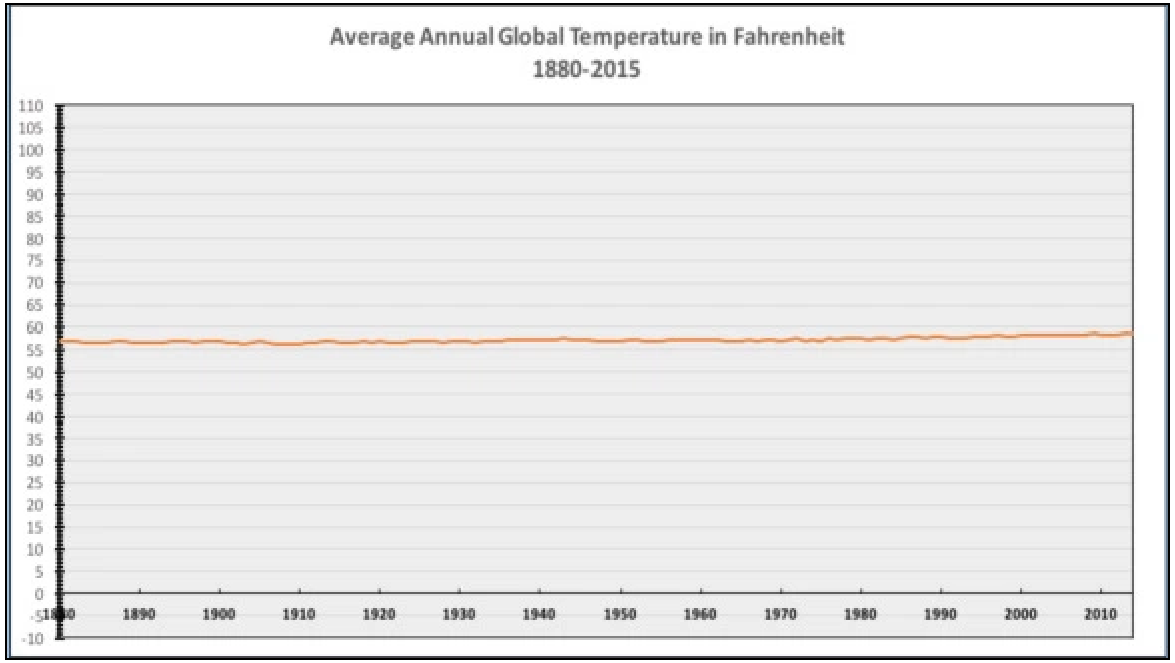
# 2. Line graph axes **need not** include zero



More California Households Have All Parents Working, Making Access to Child Care an Important Priority

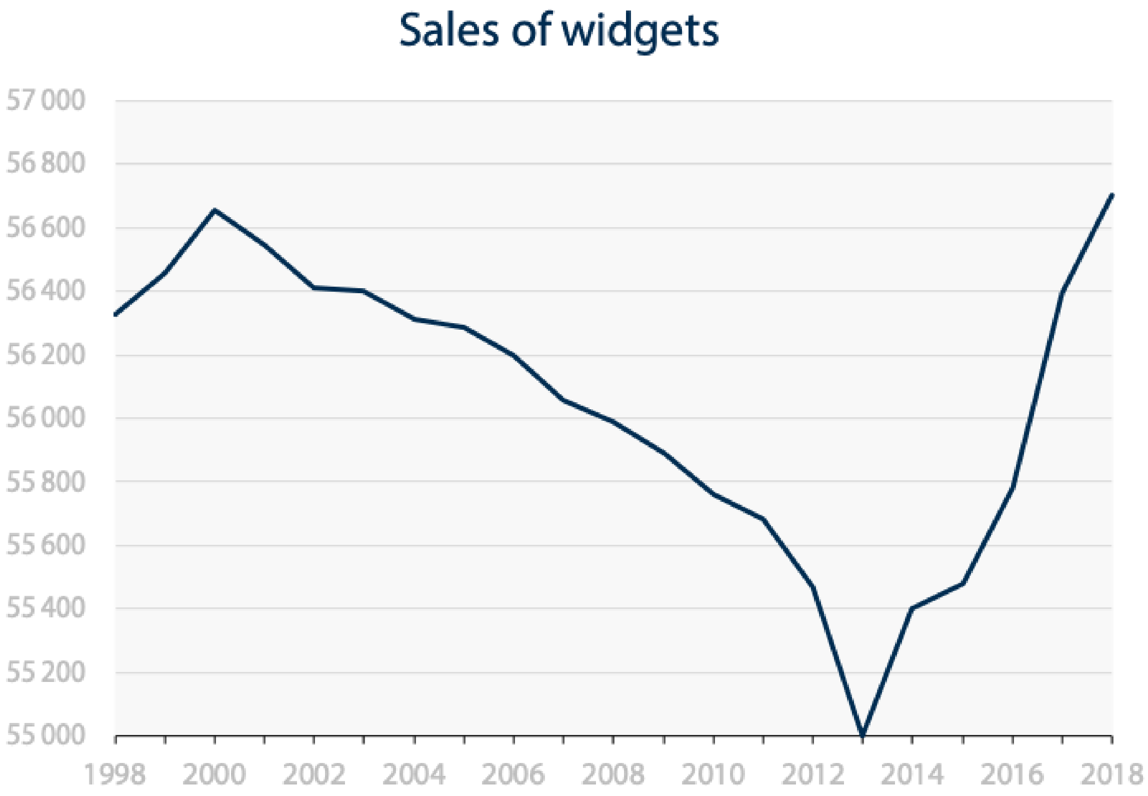
Percentage of California Households Where All Parents Work, 1970 to 2014

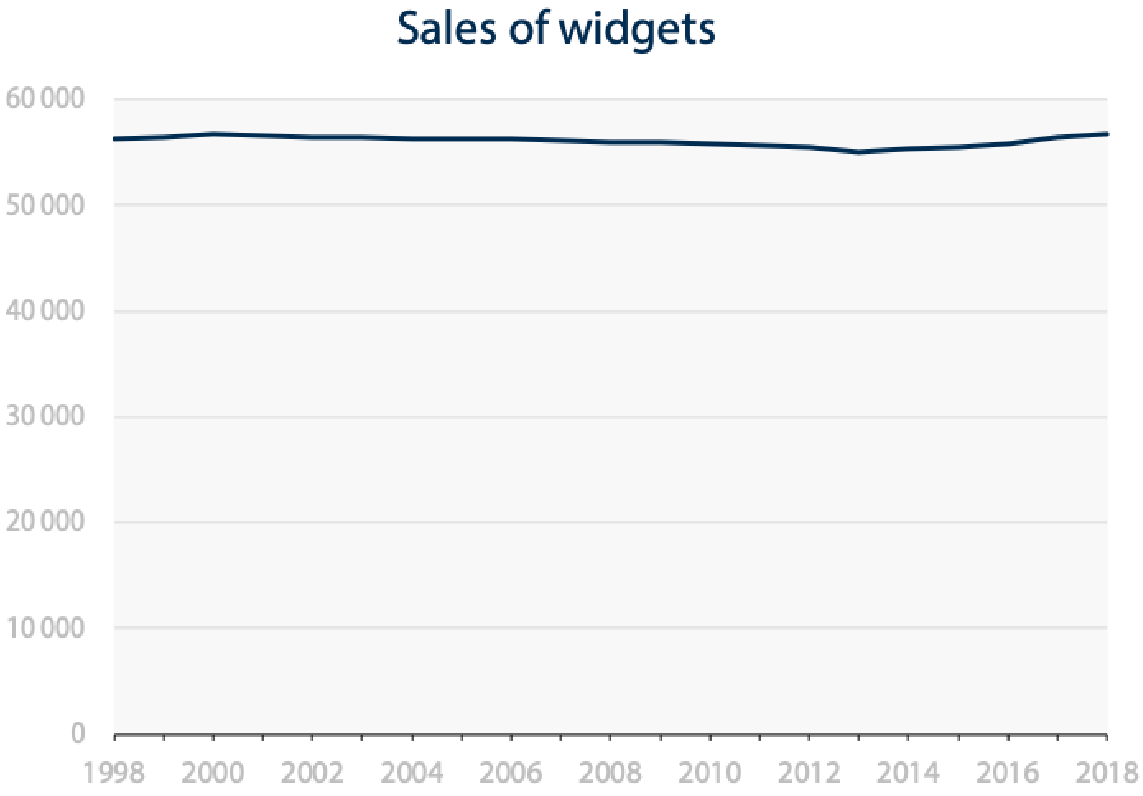
Note: A "household where all parents work" includes single-parent households and dual-earner households. Parents include stepparents and adoptive parents.
Source: Budget Center analysis of US Census Bureau data



Average Annual Global Temperature in Fahrenheit
1880-2015

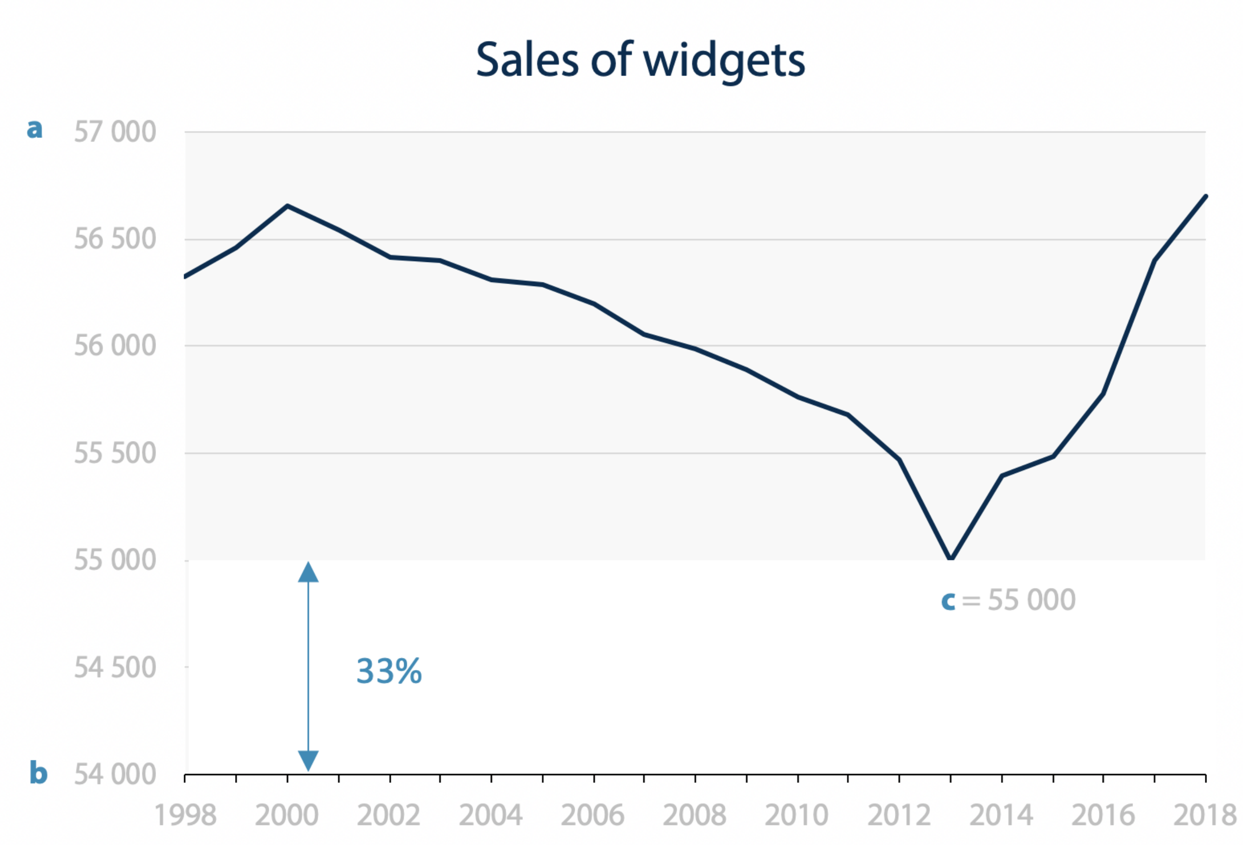
# Well…


Sales of widgets
57 000
56 800
56 600
56 400
56 200
56 000
55 800
55 600
55 400
55 200
55 000
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018


Sales of widgets
60 000
50 000
40 000
30 000
20 000
10 000
0
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018

# Solution

Empty space = $\dfrac{\text{Minimum data value - Minimum scale value}}{\text{Maximum scale value - Minimum scale value}}$



$$x = \frac{(c-b)}{(a-b)}$$

$$b = \frac{3c - a}{2}$$

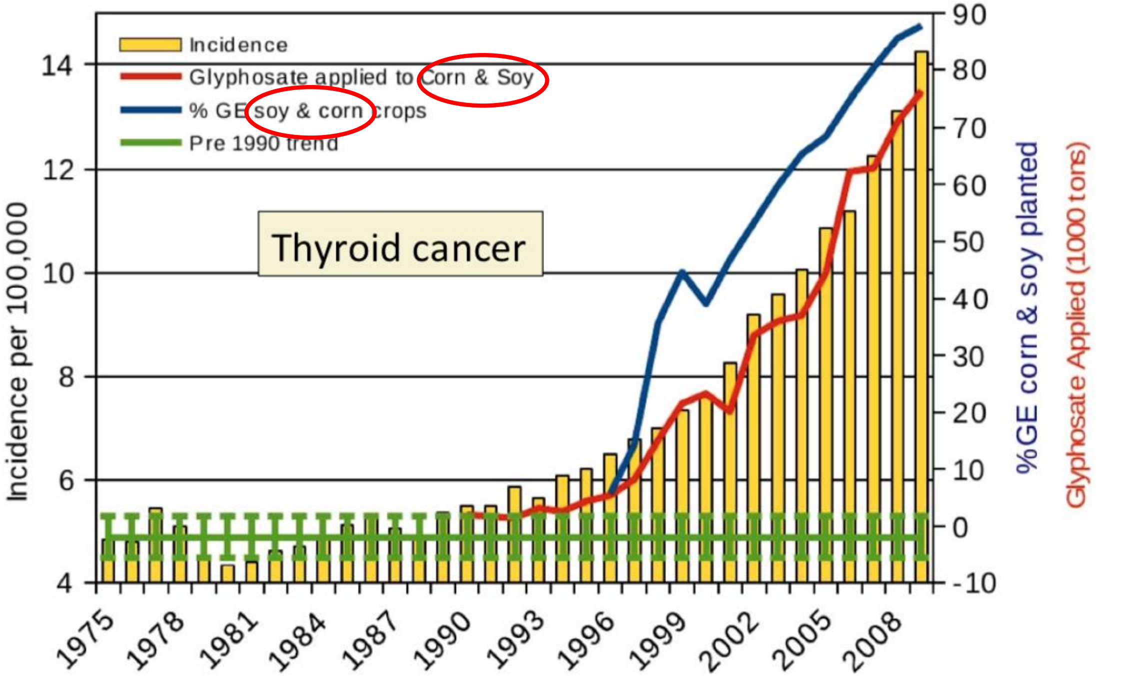
$$b = \frac{3 * 55{,}000 - 57000}{2} = 54\,000$$

Gagnon, F. (2018). A golden ratio for line charts with truncated y-axis

# 3. **No** multiple axes on a single graph

Thyroid Cancer Incidence Rate (age adjusted)

plotted against glyphosate applied to U.S. corn & soy (R = 0.988, p <= 7.612e-09)
along with %GE corn & soy crops R = 0.9377, p <= 2.152e-05
sources: USDA:NASS; SEER

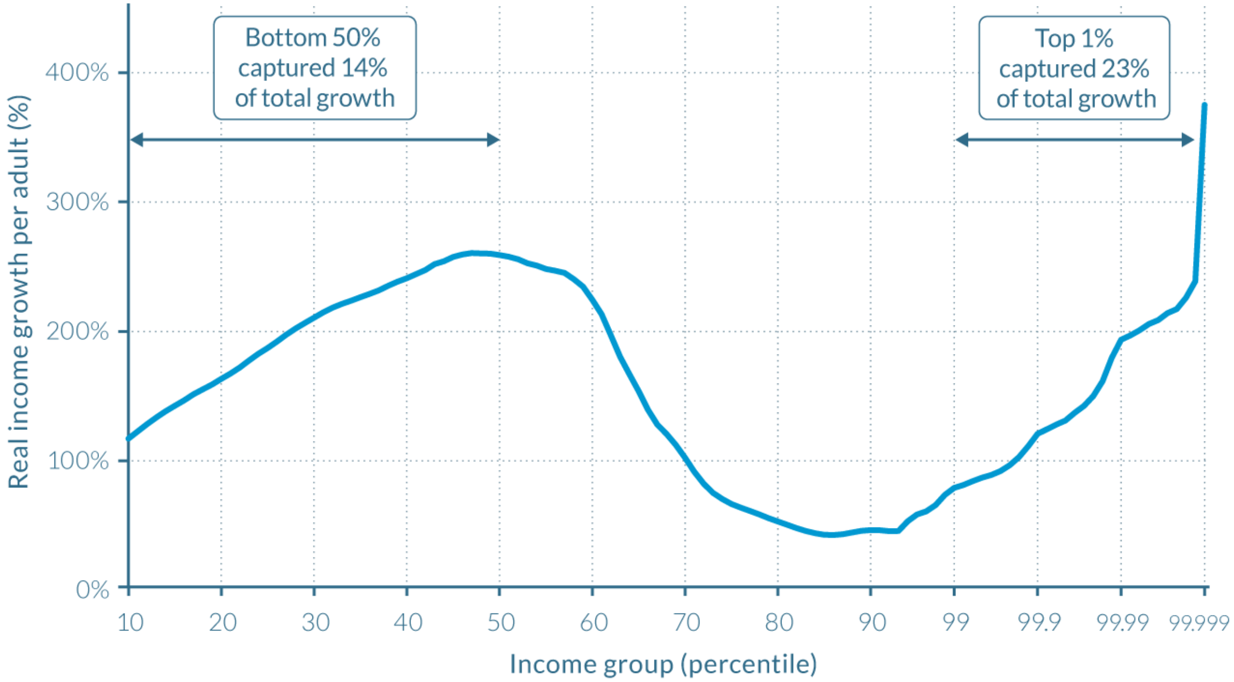


Thyroid cancer

*Figure 10, Swanson et al. Journal of Organic Systems 2014; 9(2):6-37.

How to solve this?

# 4. An axis should **not** change scales midstream

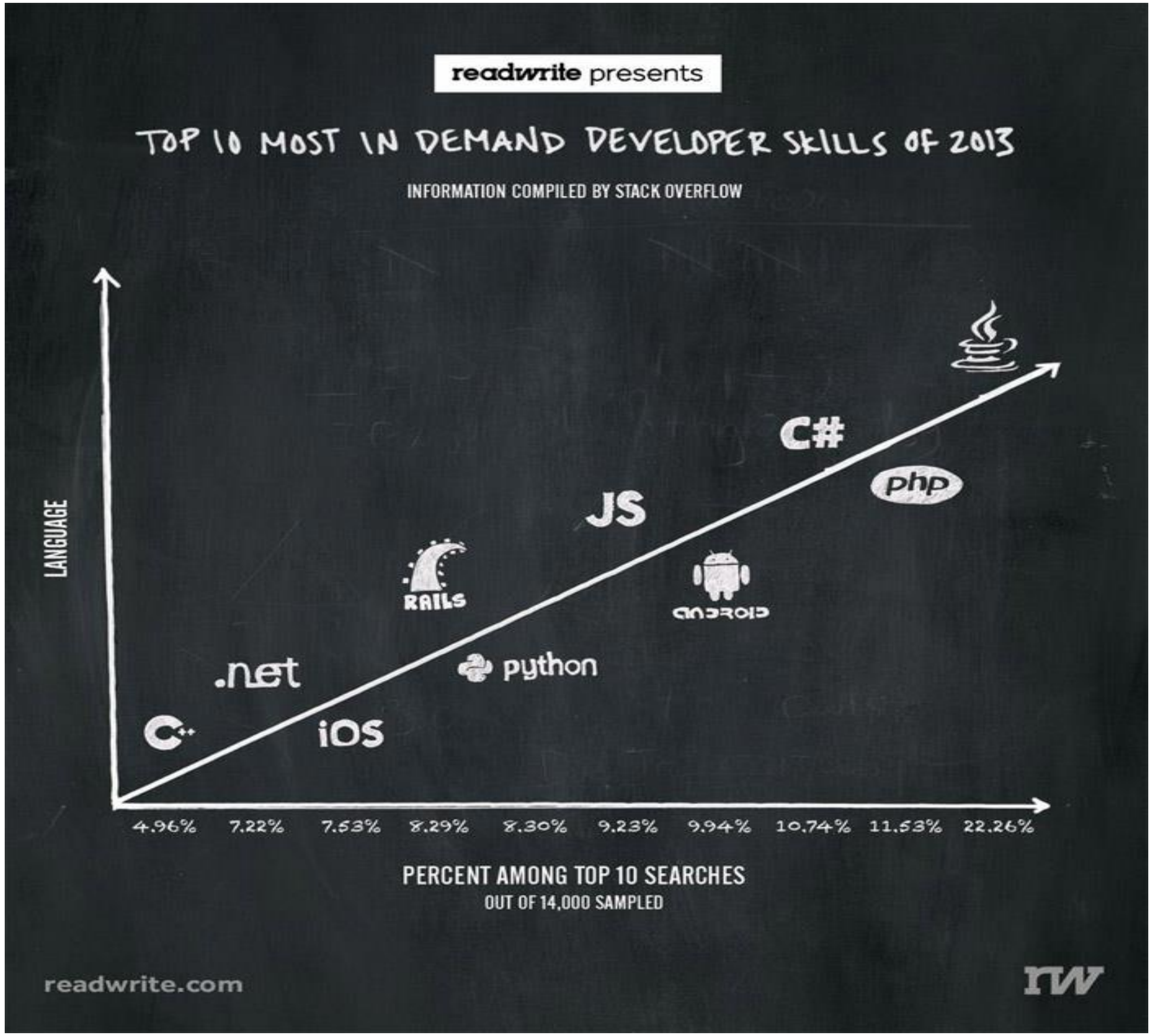


Source: WID.world (2017). See wir2018.wid.world/methodology.html for data series and notes.

On the horizontal axis, the world population is divided into a hundred groups of equal population size and sorted in ascending order from left to right, according to each group's income level. The Top 1% group is divided into ten groups, the richest of these groups is also divided into ten groups, and the very top group is again divided into ten groups of equal population size. The vertical axis shows the total income growth of an average individual in each group between 1980 and 2016. For percentile group p99p99.1 (the poorest 10% among the world's richest 1%), growth was 77% between 1980 and 2016. The Top 1% captured 23% of total growth over this period. Income estimates account for differences in the cost of living between countries. Values are net of inflation.

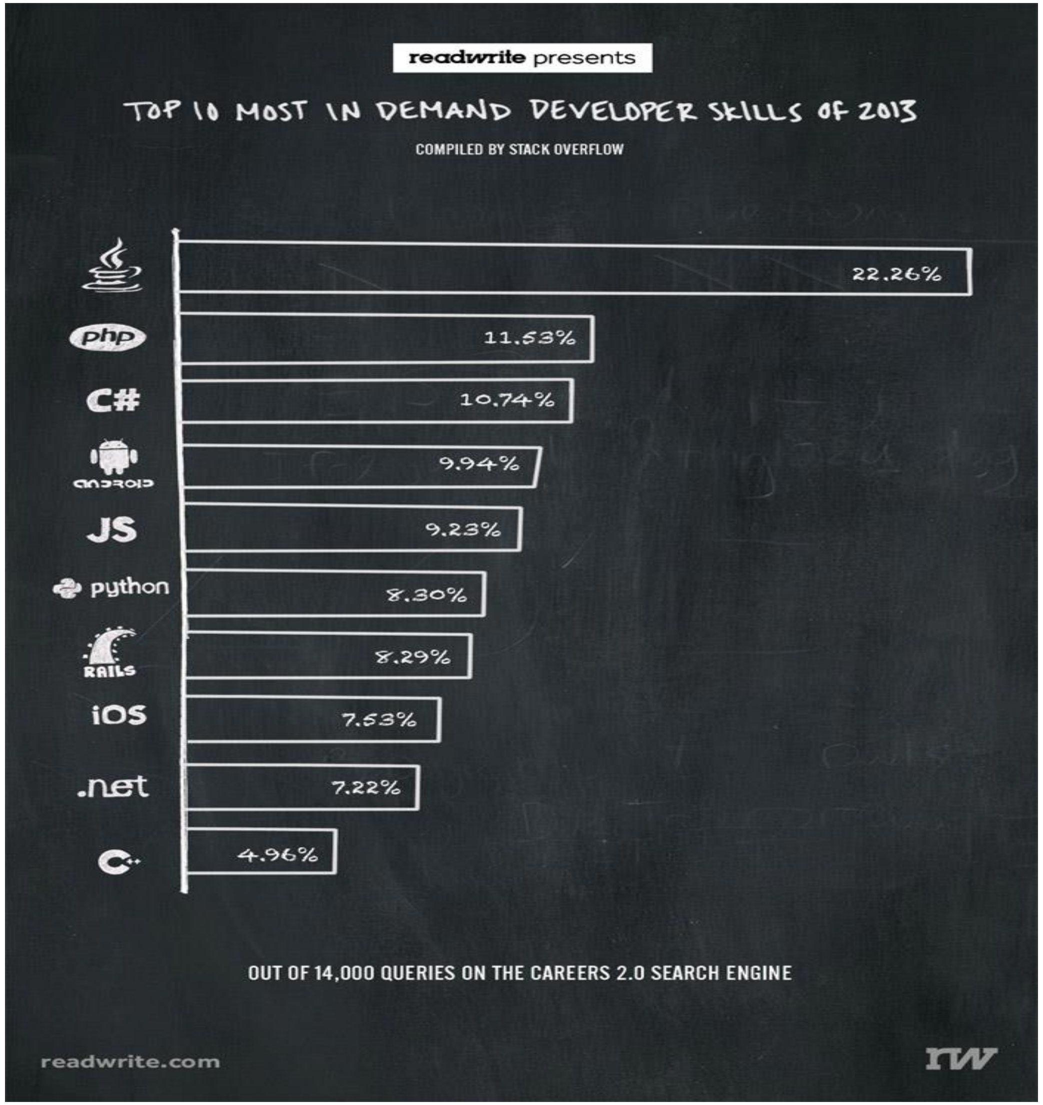
**Use log scales wisely!**

# 5. An axis should have *something* on it



# Like this

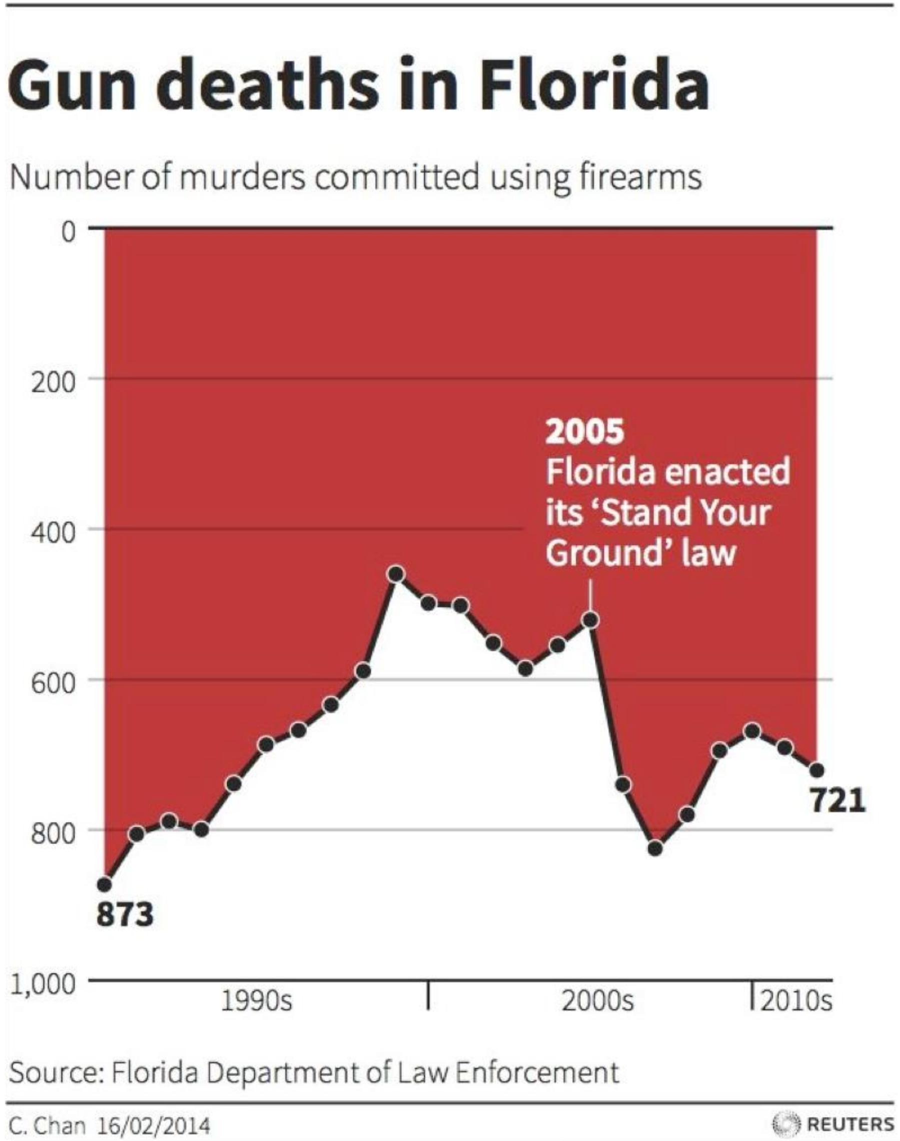
Tools such as WebPlotDigitizer can help you get the missing info



# 6. **Don’t** invert the axis(?)

The author defended it!

What’s your stance?



Gun deaths in Florida

Number of murders committed using firearms

Source: Florida Department of Law Enforcement

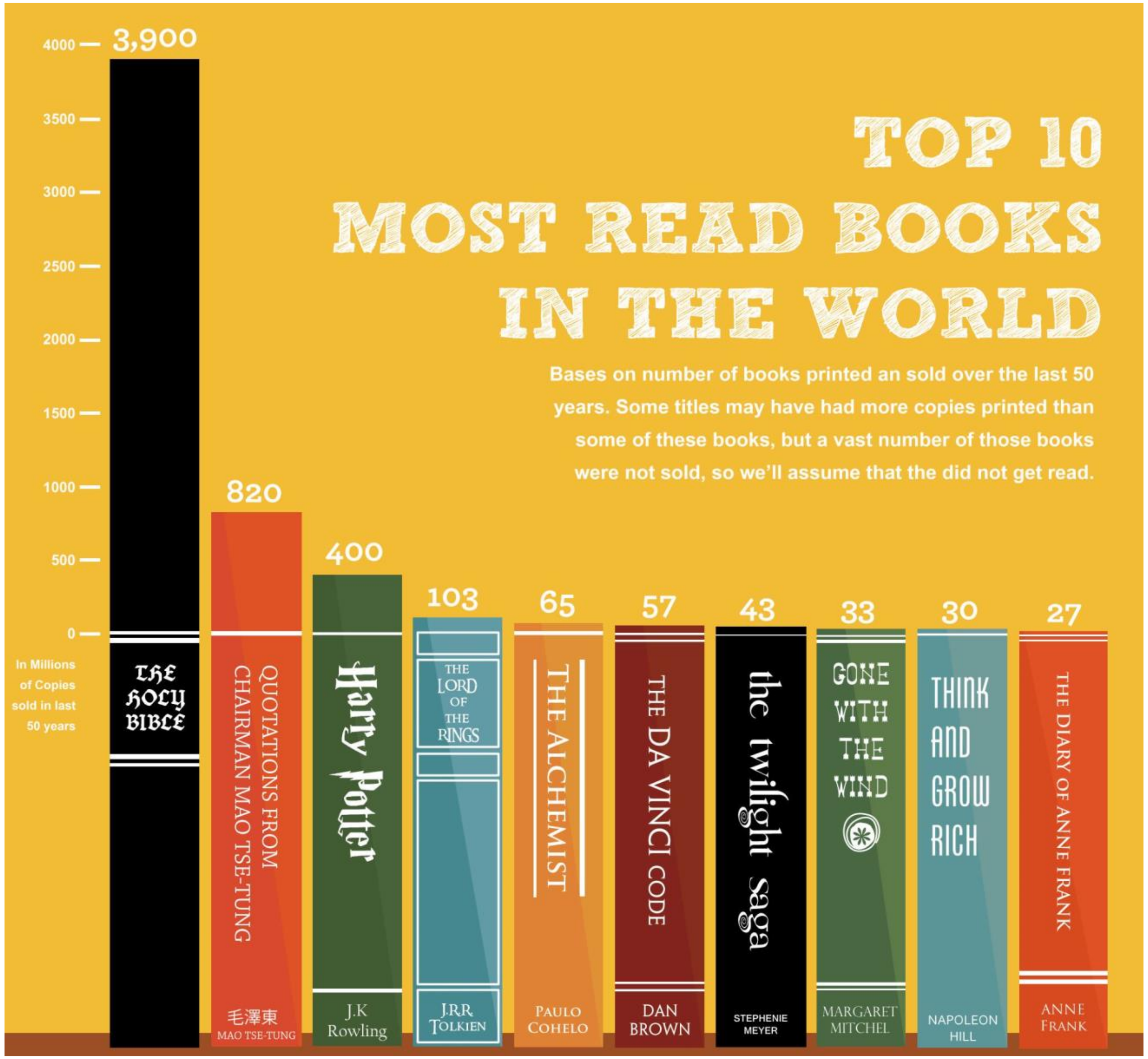
C. Chan 16/02/2014

REUTERS

# Proportional Ink

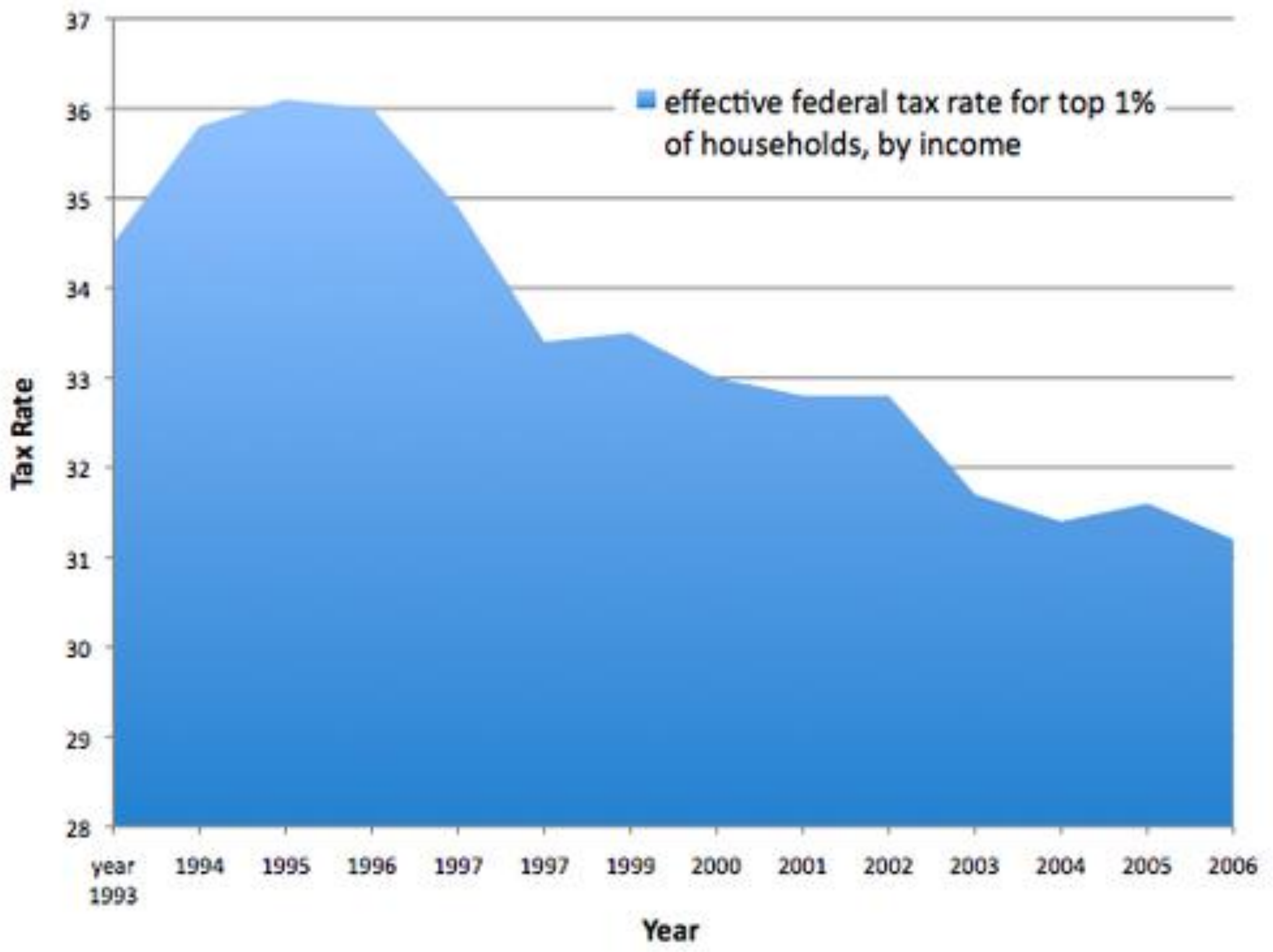
*“When a shaded region is used to represent a numerical value, the area of that shaded region should be directly proportional to the corresponding value”*

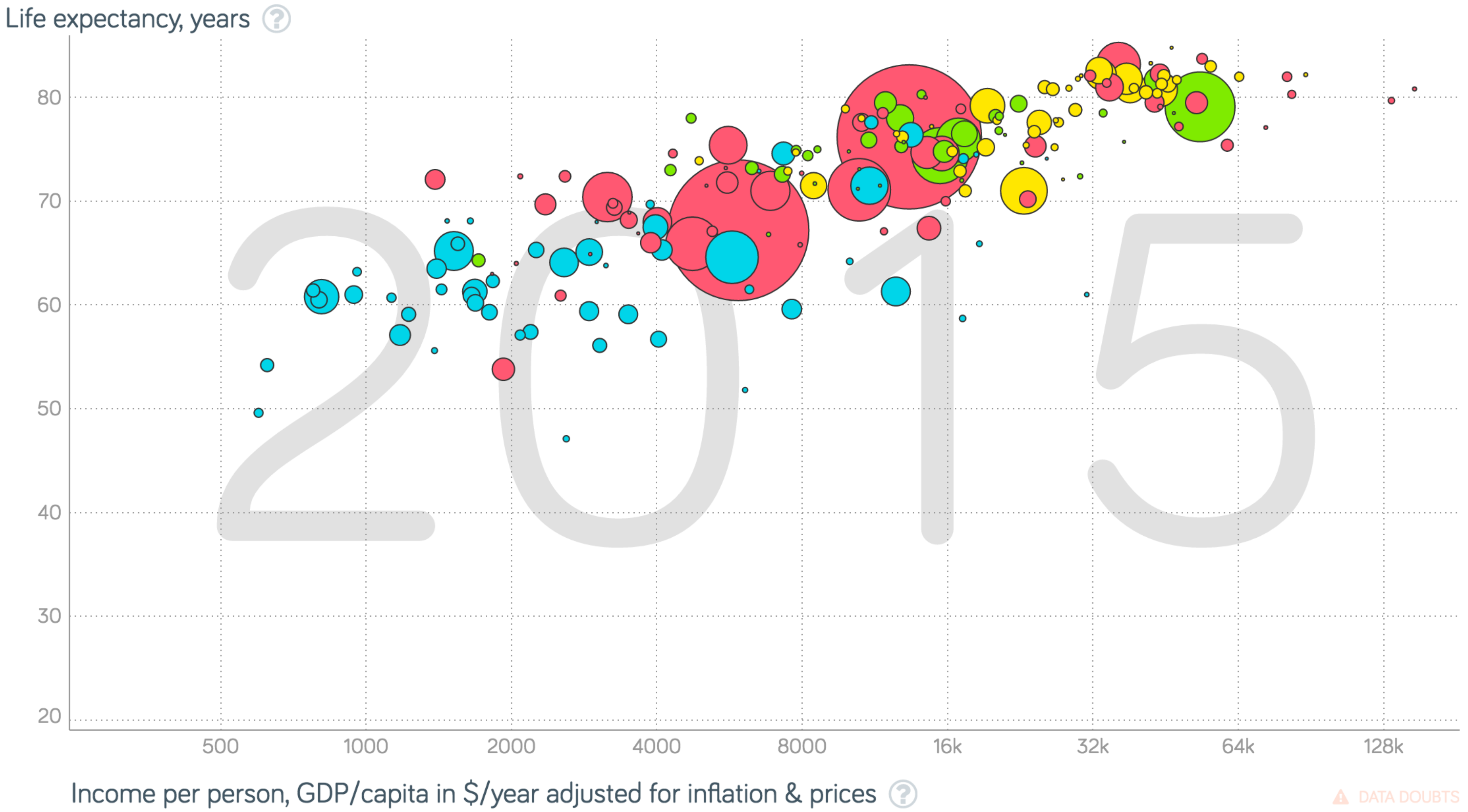
Extends the argument for misleading axes



# TOP 10 MOST READ BOOKS IN THE WORLD

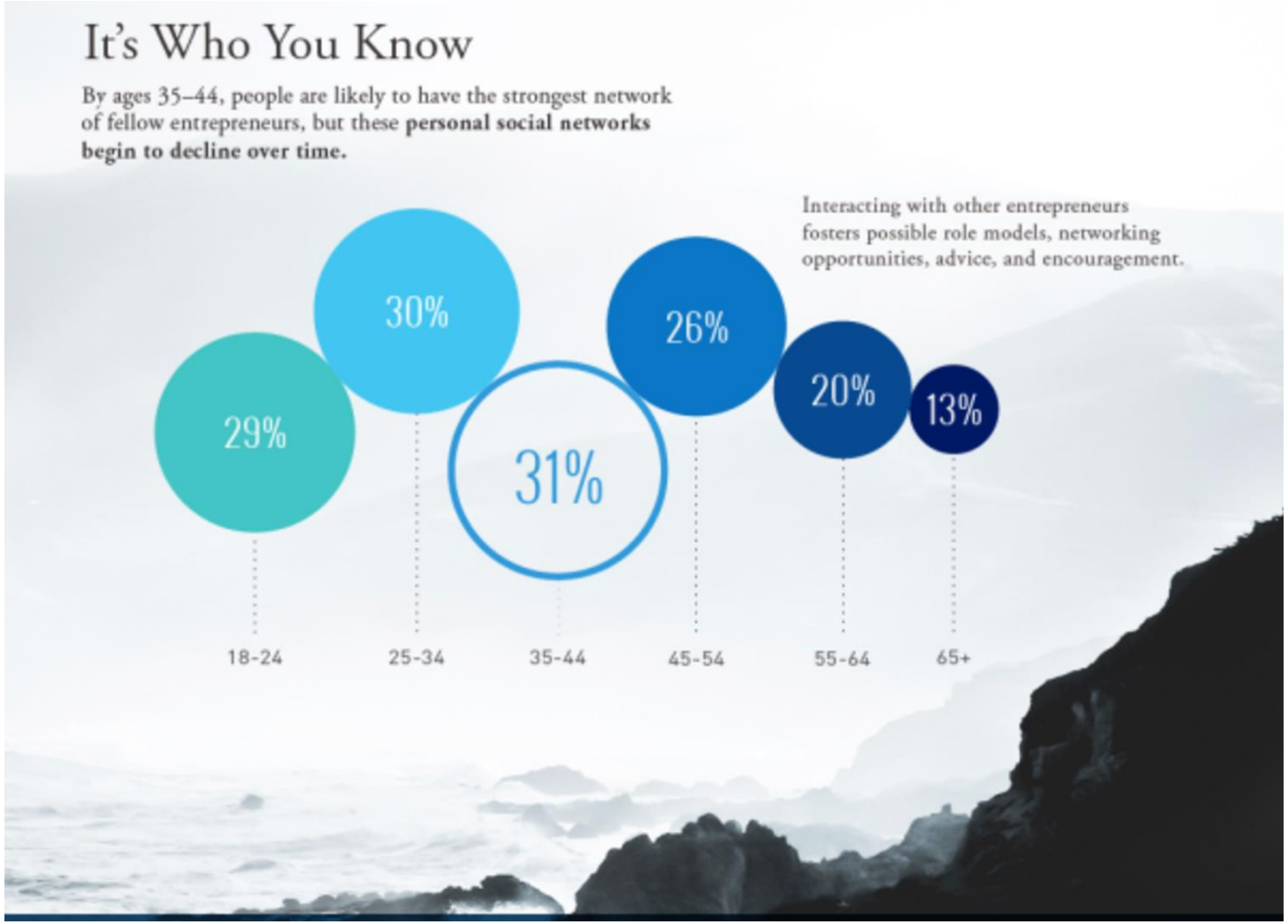
Bases on number of books printed an sold over the last 50 years. Some titles may have had more copies printed than some of these books, but a vast number of those books were not sold, so we'll assume that the did not get read.


🟦 effective federal tax rate for top 1% of households, by income
Tax Rate
37
36
35
34
33
32
31
30
29
28
year 1993
1994
1995
1996
1997
1997
1999
2000
2001
2002
2003
2004
2005
2006
Year



**Should we violate the principle?**

**Radius ≠ Area**

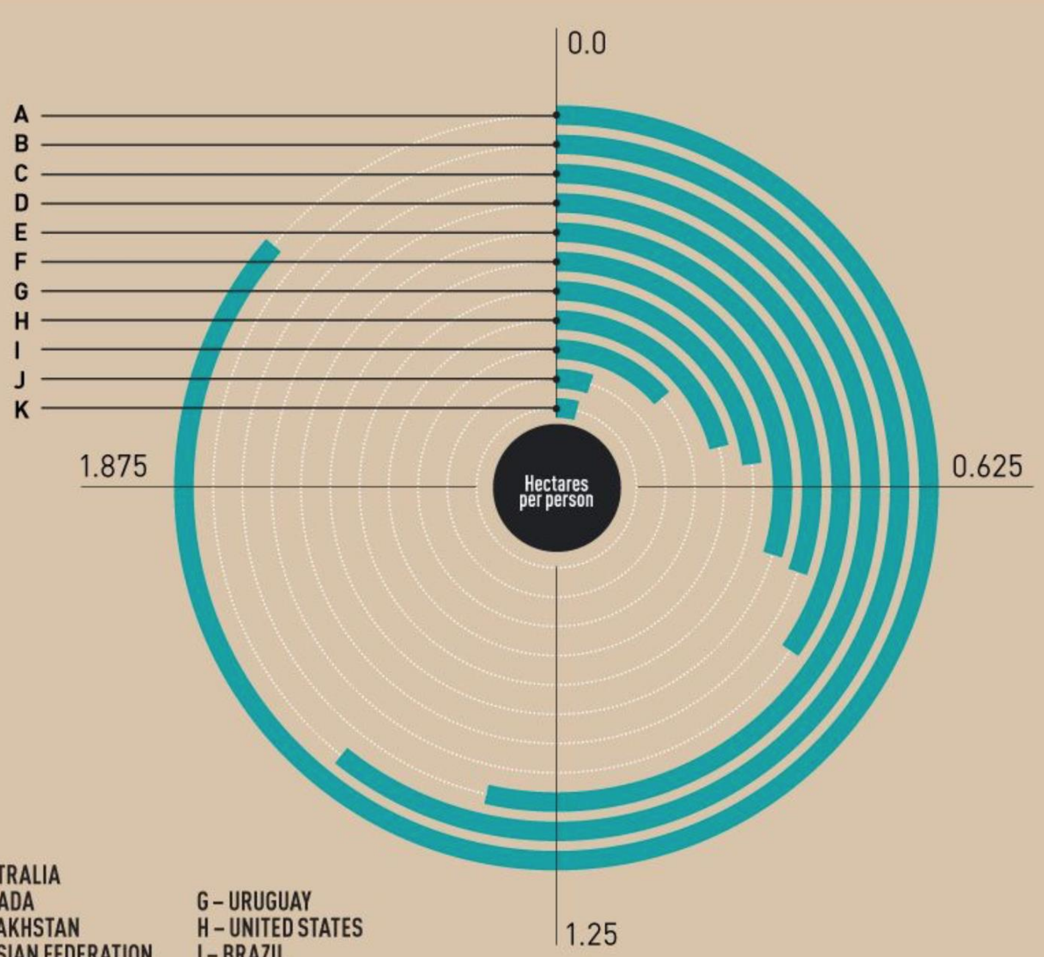


## It's Who You Know

By ages 35–44, people are likely to have the strongest network of fellow entrepreneurs, but these **personal social networks begin to decline over time.**

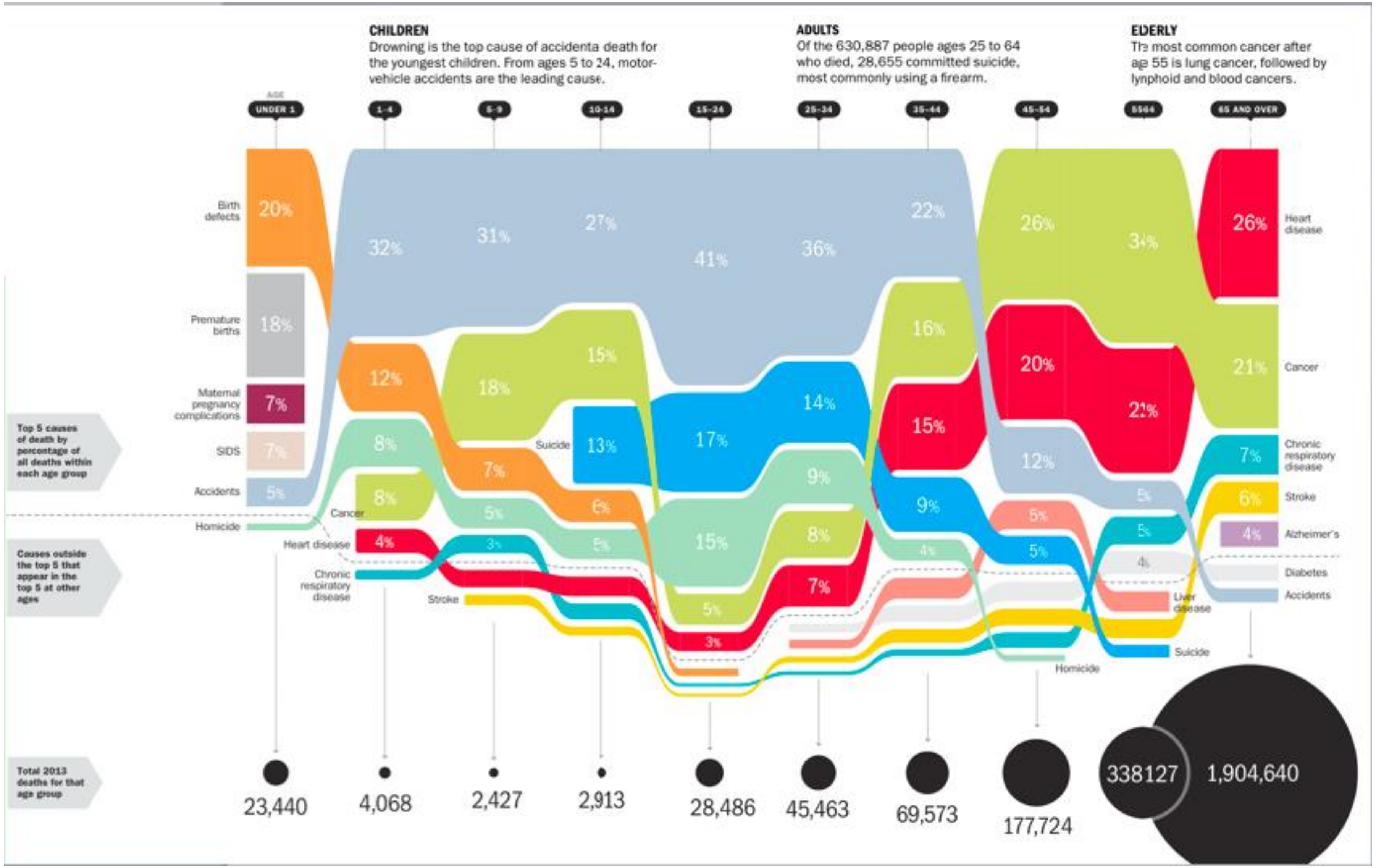
Interacting with other entrepreneurs fosters possible role models, networking opportunities, advice, and encouragement.

# ARABLE LAND PER CAPITA (HECTARES PER PERSON)



A – AUSTRALIA
B – CANADA
C – KAZAKHSTAN
D – RUSSIAN FEDERATION
E – ARGENTINA
F – UKRAINE
G – URUGUAY
H – UNITED STATES
I – BRAZIL
J – NEW ZEALAND
K – CHILE

Source:
*CIA Factbook 2012* and FAO 2011



**CHILDREN**
Drowning is the top cause of accidenta death for the youngest children. From ages 5 to 24, motor-vehicle accidents are the leading cause.

**ADULTS**
Of the 630,887 people ages 25 to 64 who died, 28,655 committed suicide, most commonly using a firearm.

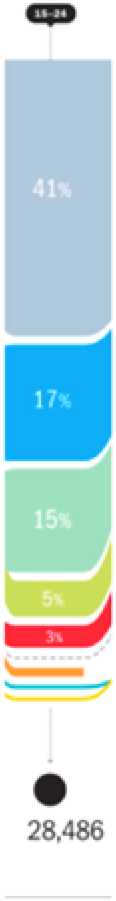
**EIDERLY**
Tl϶ most common cancer after ag϶ 55 is lung cancer, followed by lymphoid and blood cancers.

Top 5 causes of death by percentage of all deaths within each age group

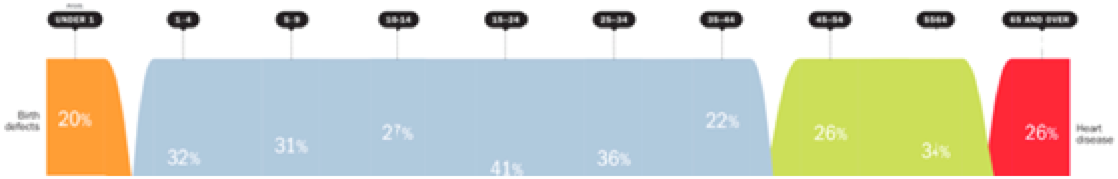
Causes outside the top 5 that appear in the top 5 at other ages

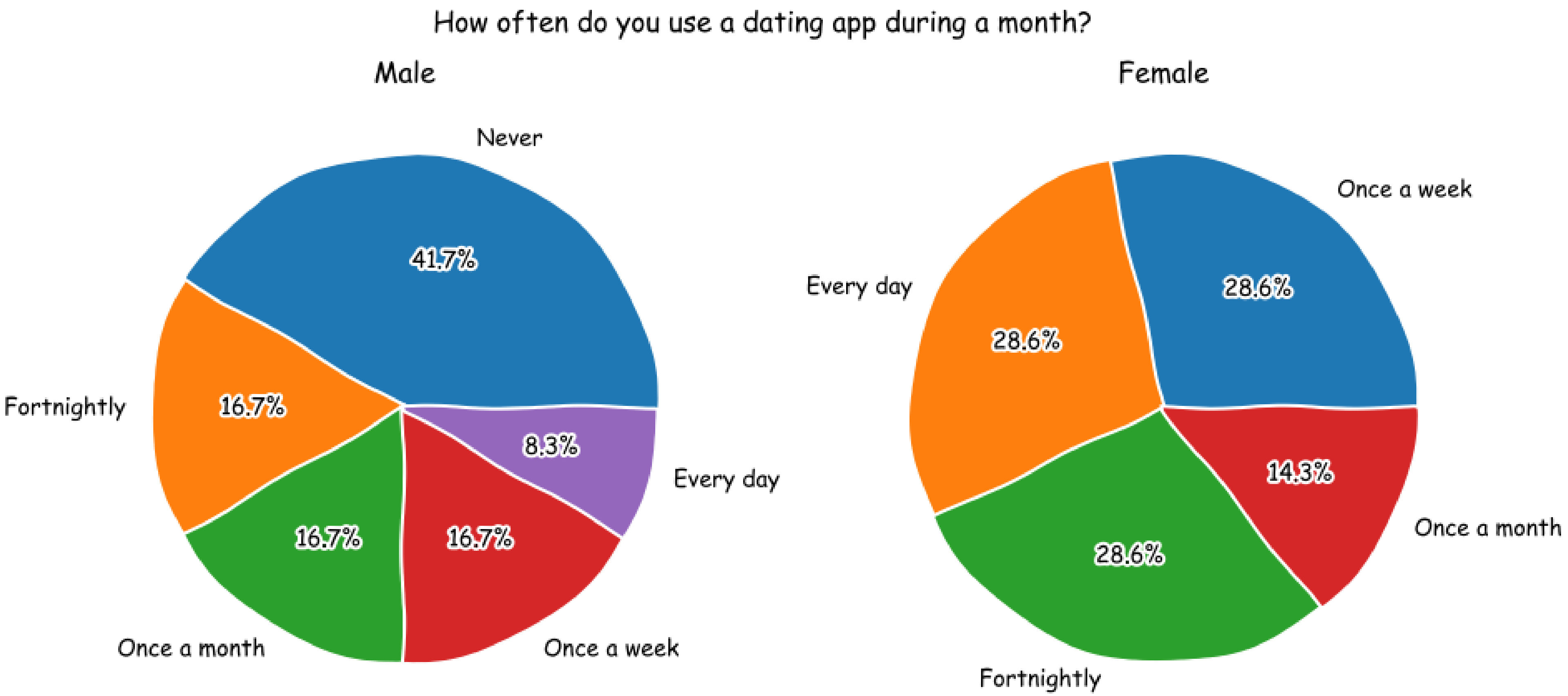
Total 2013 deaths for that age group

Along vertical slices, ink is proportional to value because shaded areas represent the fraction of a fixed number of deaths (here 28,486) from each cause.



Along horizontal slices, ink is not proportional because total deaths differ widely by age group. Far more people 65 and older died of heart disease (red) than children age1-4 die of accidents (blue-gray), but the latter takes more ink because it represents a larger percentage of the (relatively few) total deaths at that age.





**What is wrong in this one?**

# Perspective

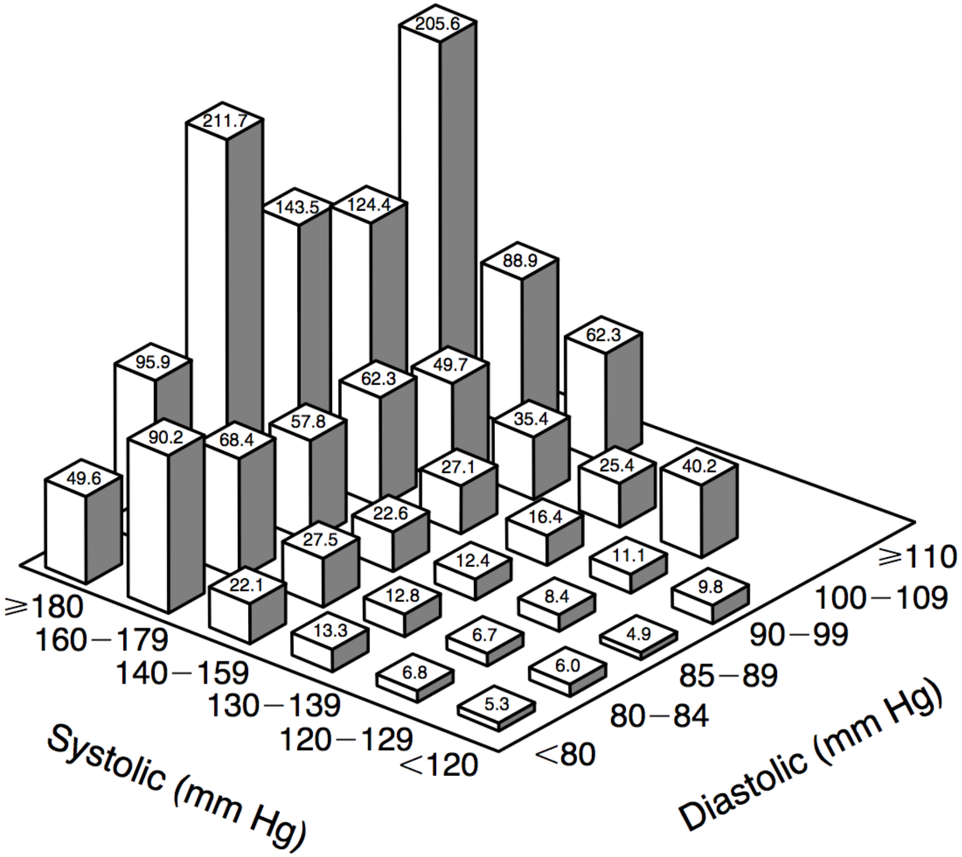
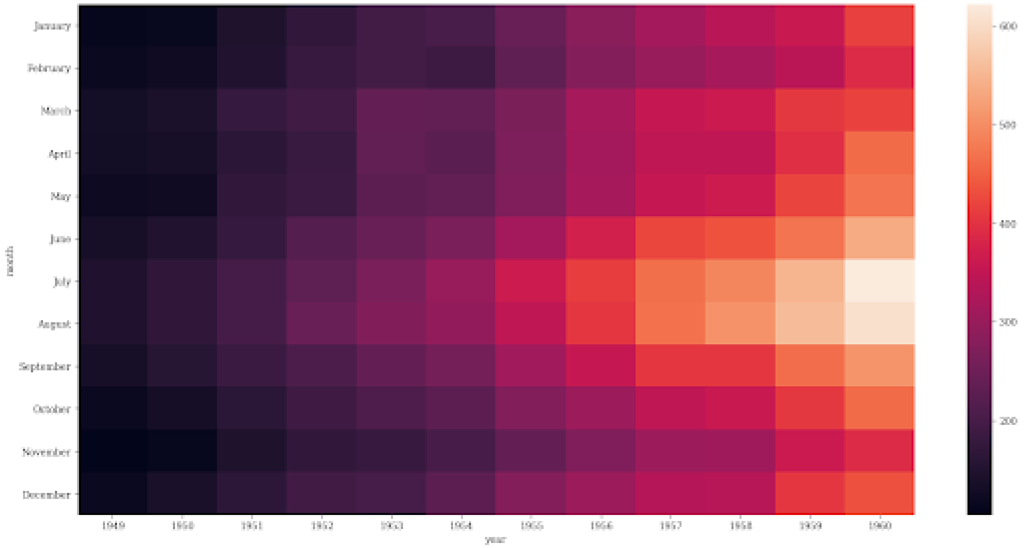
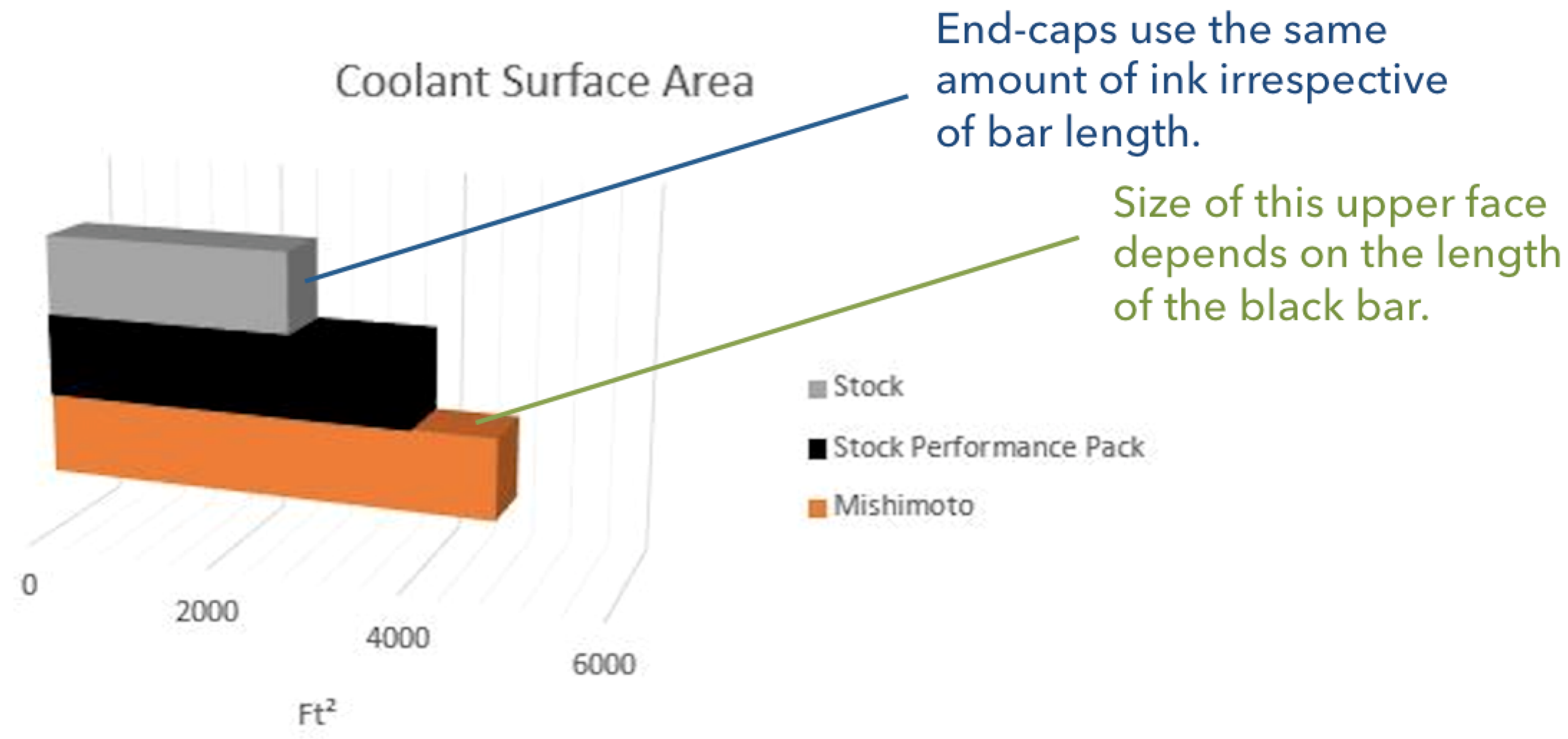


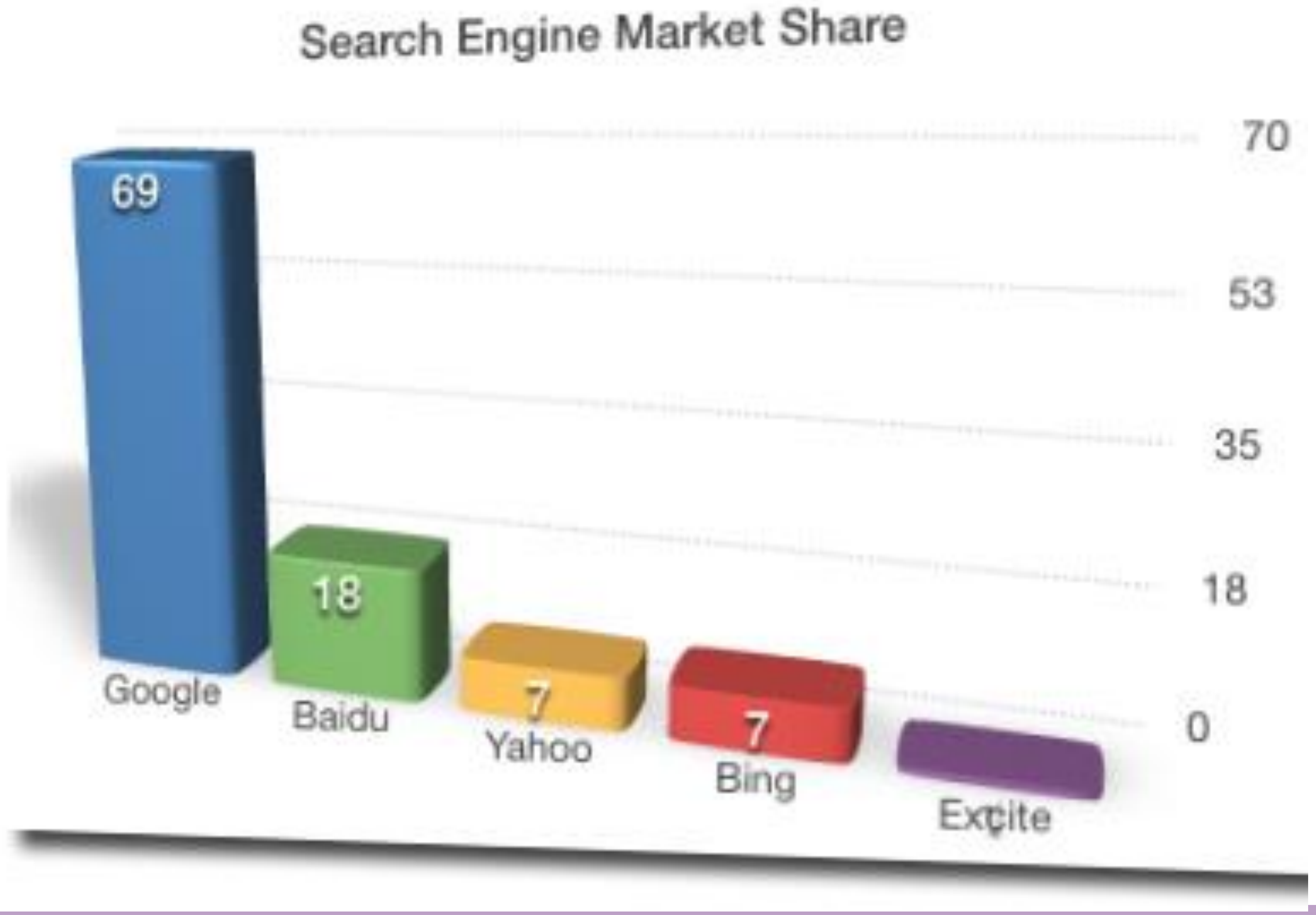
Figure 2. Age-Adjusted Rate of End-Stage Renal Disease Due to Any Cause per 100,000 Person-Years, According to Systolic and Diastolic Blood Pressure in 332,544 Men Screened for MRFIT.

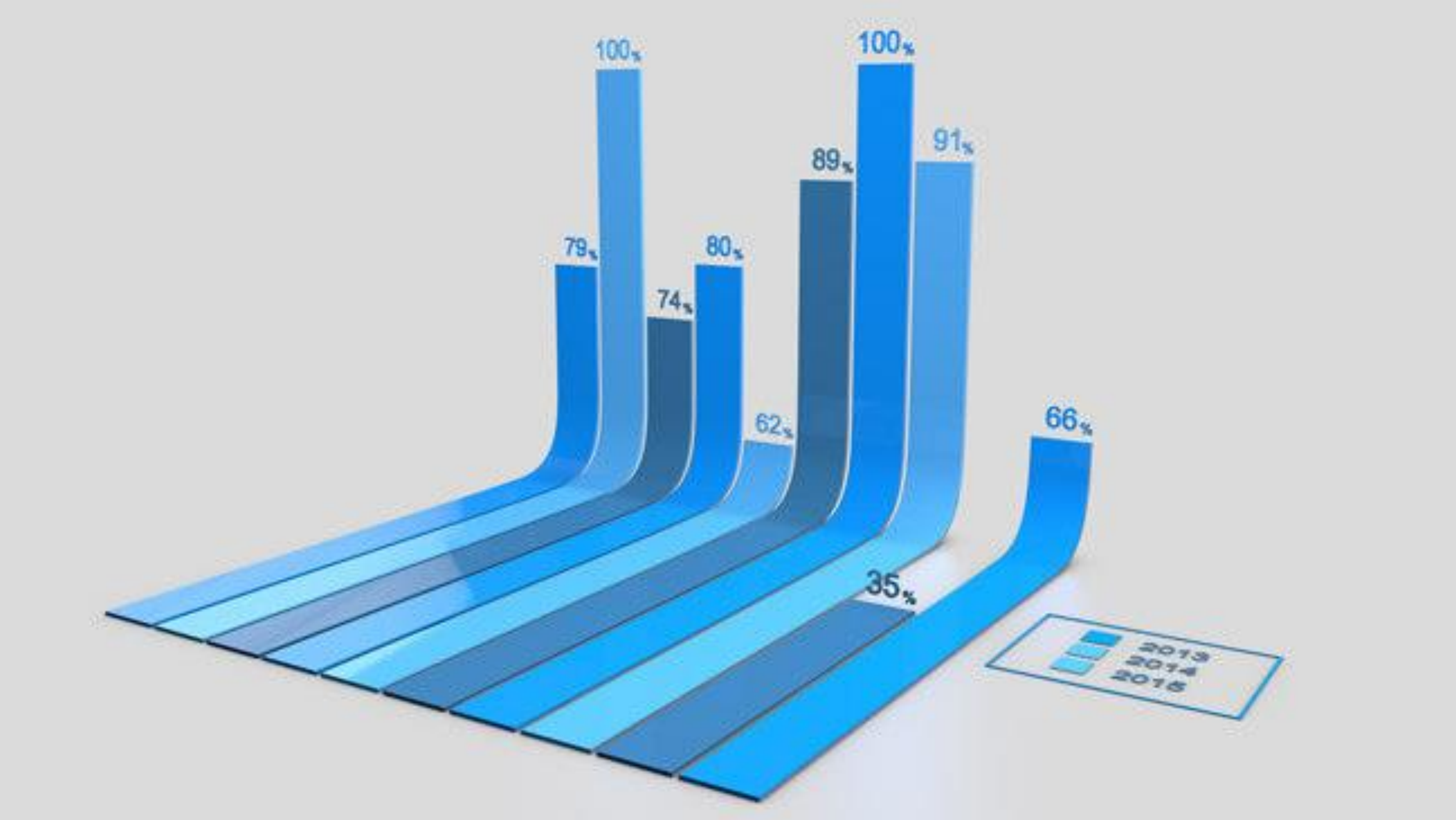


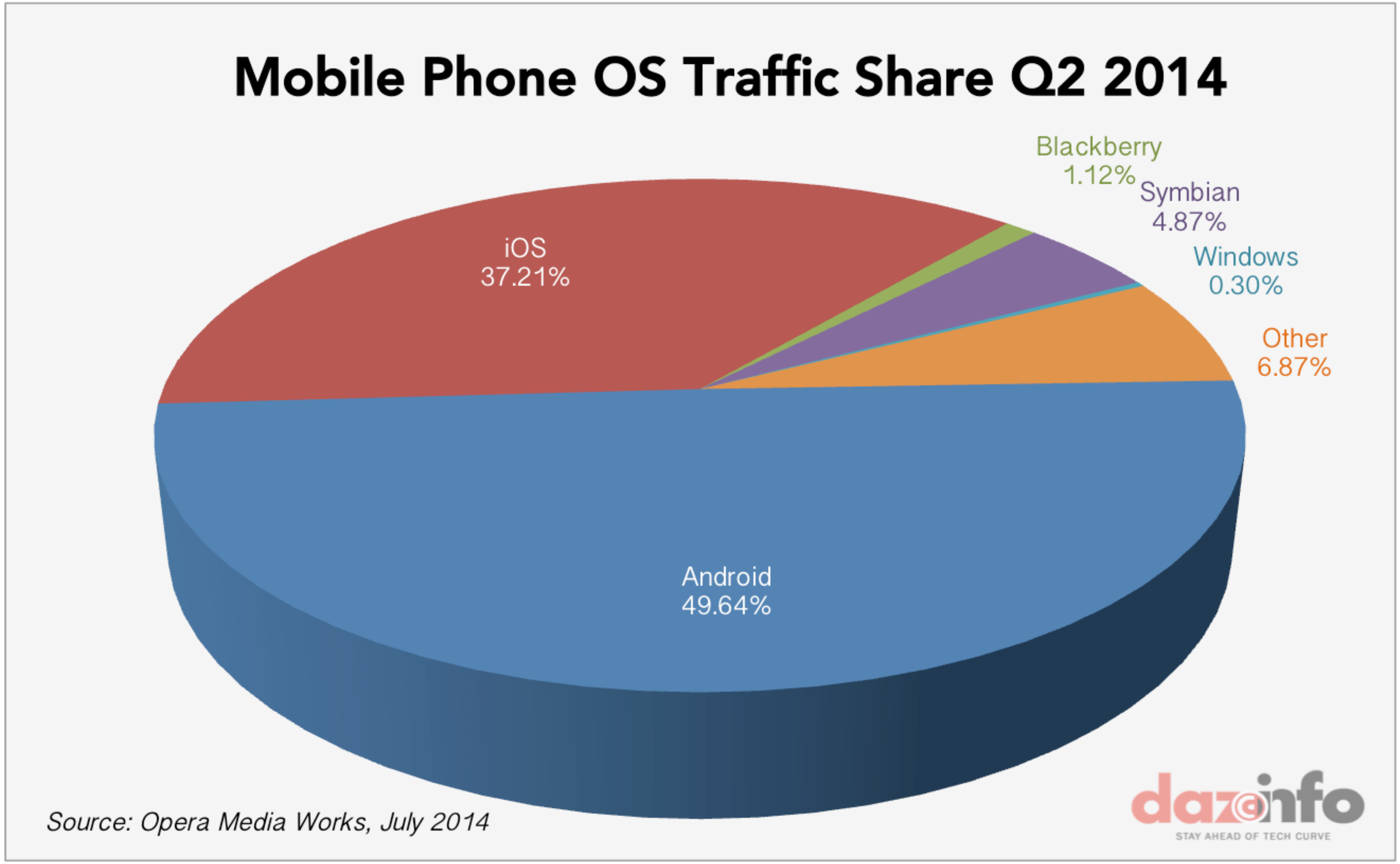


End-caps use the same amount of ink irrespective of bar length.

Size of this upper face depends on the length of the black bar.


Search Engine Market Share
70
53
35
18
0
69
18
7
7
Google
Baidu
Yahoo
Bing
Excite

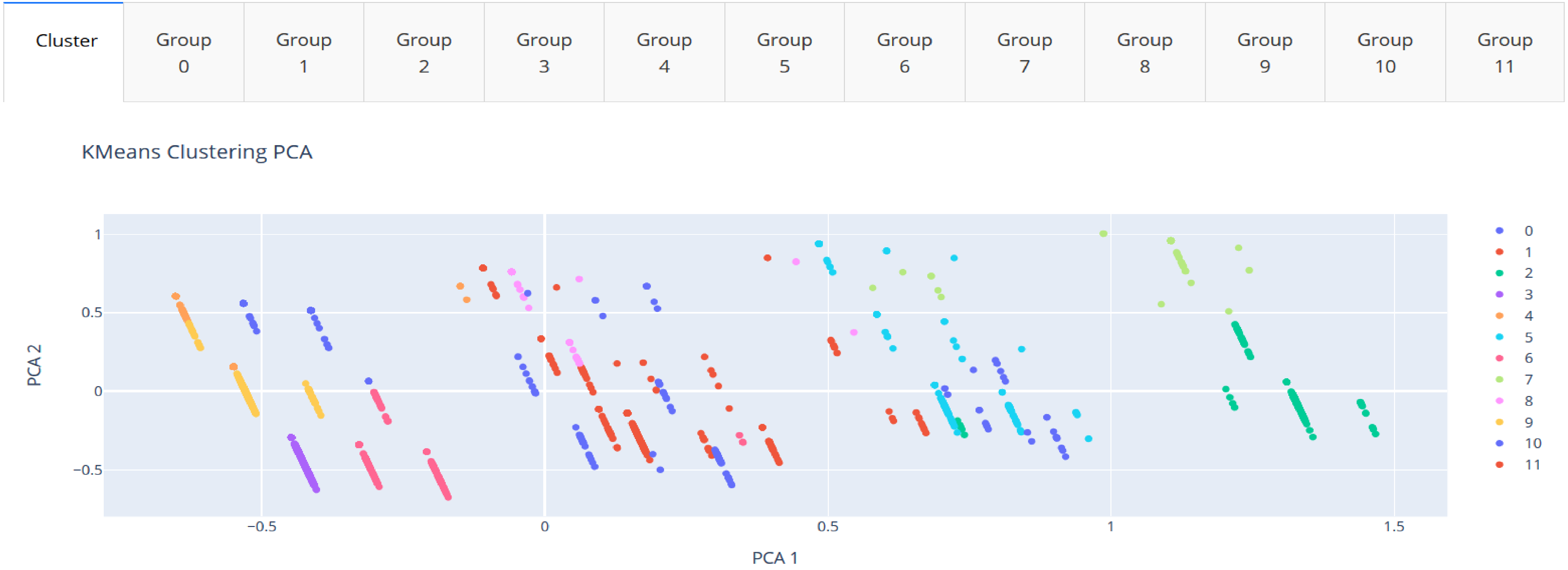

100%
79%
74%
80%
62%
89%
100%
91%
35%
66%
2013
2014
2015



Mobile Phone OS Traffic Share Q2 2014

*Source: Opera Media Works, July 2014*

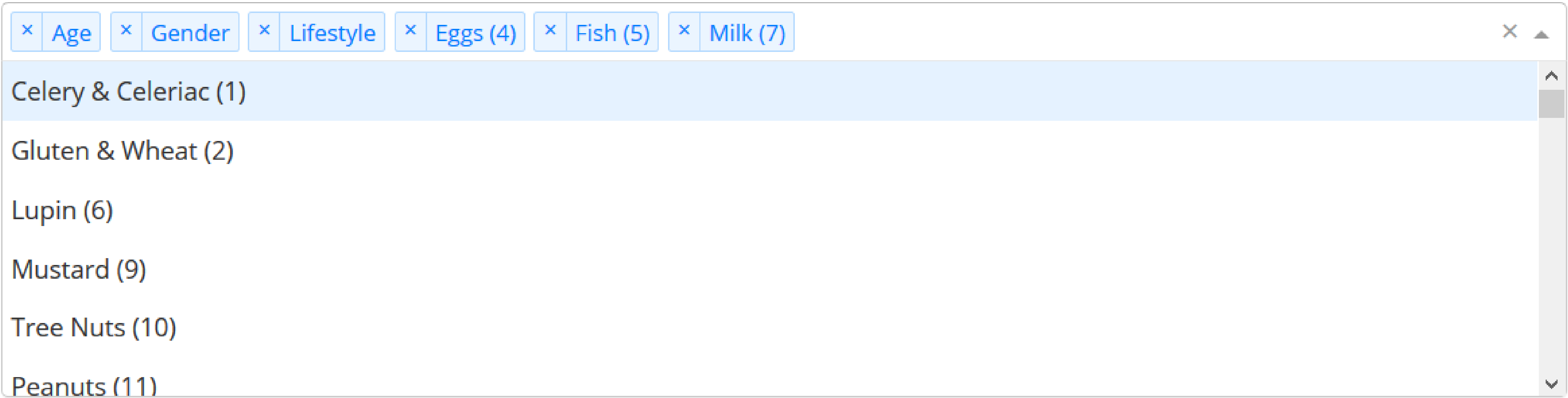
# The most hated visualisation of all time


Cluster
Group 0
Group 1
Group 2
Group 3
Group 4
Group 5
Group 6
Group 7
Group 8
Group 9
Group 10
Group 11
KMeans Clustering PCA
PCA 2
1
0.5
0
−0.5
PCA 1
−0.5
0
0.5
1
1.5
● 0
● 1
● 2
● 3
● 4
● 5
● 6
● 7
● 8
● 9
● 10
● 11

# Clustering

- The **IDEA** of this plot was to show stakeholders (i.e. the people from the listed restaurants looking at this dashboard) that users/profiles can be grouped in NEW WAYS
- Therefore, we process the user/profile data with a method called *k-means*
- In short, this method analyses the **FEATURES** of all users/profiles and groups them into clusters

# Step 1: Select the Features



× Age × Gender × Lifestyle × Eggs (4) × Fish (5) × Milk (7)

Celery & Celeriac (1)

Gluten & Wheat (2)

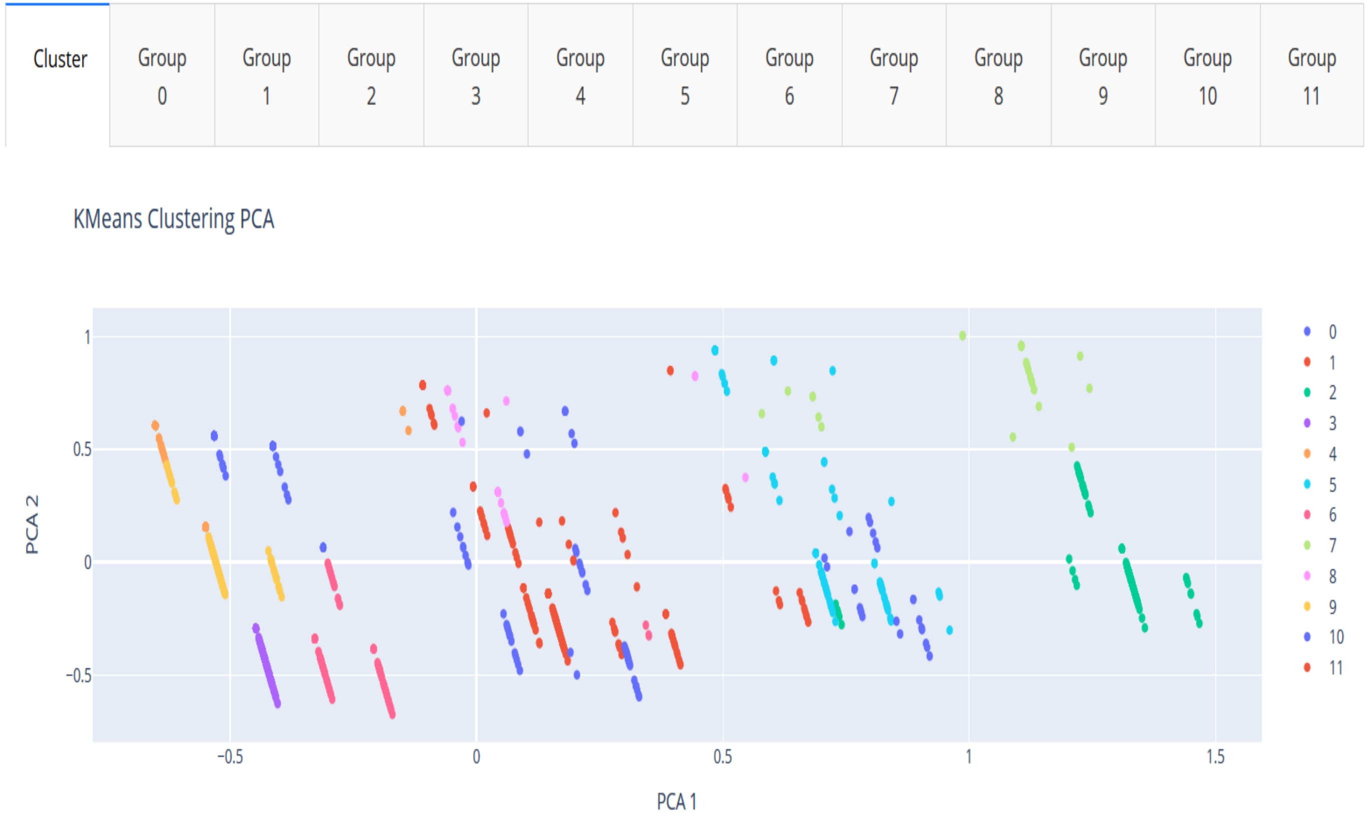
Lupin (6)

Mustard (9)

Tree Nuts (10)

Peanuts (11)

# Step 2: Visualise the clustering



-In this example, the algorithm has found 12 groups using the 6 features selected in the previous slide

-Each dot represents a user/profile

-The problem is that we selected 6 features, but since each feature corresponds to one axis, it is impossible to do 6-dimension plots

-Therefore, we use a method called PCA to reduce the dimensionality from 6 to 2

# Step 3: Understand each cluster

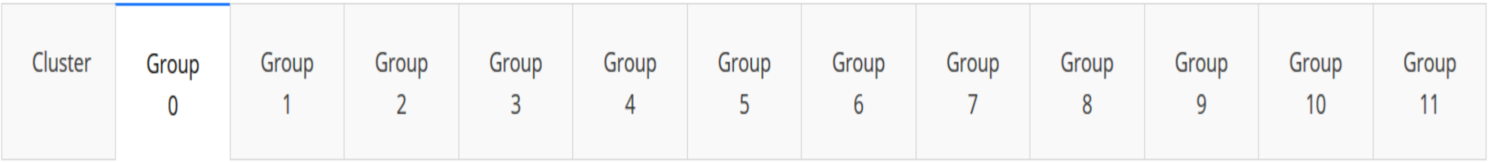
-For each group found, you have a tab at the top

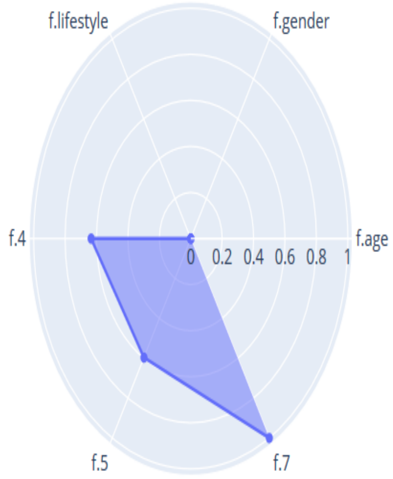
-This tab takes you to a radial chart, in which the **<u>IDEA</u>** is to explain which are the features that influence the most to the creation of a certain cluster

-In this case, we can see that for group 0 it was feature 7 (milk allergy) the one that most defines this group

-This can mean that either it is highly likely that most people in this group are either <u>all allergic</u> or <u>not allergic</u> to milk

-However, at this stage we <u>don't know which is it!</u>





# As for the marking…

What am I expecting? How will I grade?

# One very simple recommendation…