Data science for everyone

Prof. Jones-Rooy & Prof. Policastro

March 2, 2020

6.1: Visualizations

ANNOUNCEMENTS

Feedback on assignments

Lab 1& 2 are returned

HW 1 to be reviewed in section this week

Lab 3 & 4 will be returned ASAP

HW2 code will be returned March 10

We will be discussing questions similar to HW2 conceptual questions

1. **Assignments**
   - Lab 3 is out; due Wed., March 4, 8p
   - Homework 2 is out; due Mon., March 9, 8p
   - Lab 4 out Wed., March 4; due Mon. March 9, 8p
   - Project description released tonight; due May 4

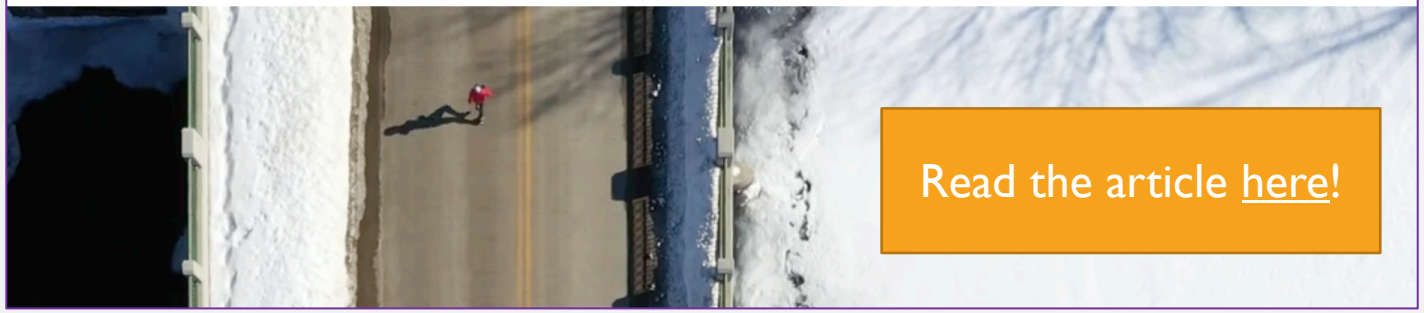
Notice earlier due date

2. **Midterm**
   - Midterm exam in lecture, Wed. March 11
   - Review in lecture, Mon., March 9
   - TAs will hold review session office hours Tuesday, March 10, 2-5p in CDS
   - No sections next week
   - Sections this week: Go over HW1, Lab 4, review (bring your Qs!)

DATA SCIENCE IN THE NEWS!

Why Are American Women Running Faster Than Ever? We Asked Them — Hundreds of Them

By Talya Minsberg and Kevin Quealy Feb. 28, 2020



Read the article here!

Selecting on the DV

More than 450 women will race in the U.S. Olympic Trials marathon in Atlanta on Saturday. That's a huge number, reflecting changes in rules, innovative shoe technology and a sea change in women's running.

Amateur women are running faster than ever, and, through communal networks online, showing others how to do so.

To qualify to race on Saturday, a woman had to complete a marathon in 2 hours 45 minutes or faster sometime in the last three years, roughly a pace of 6 minutes 17 seconds per mile.

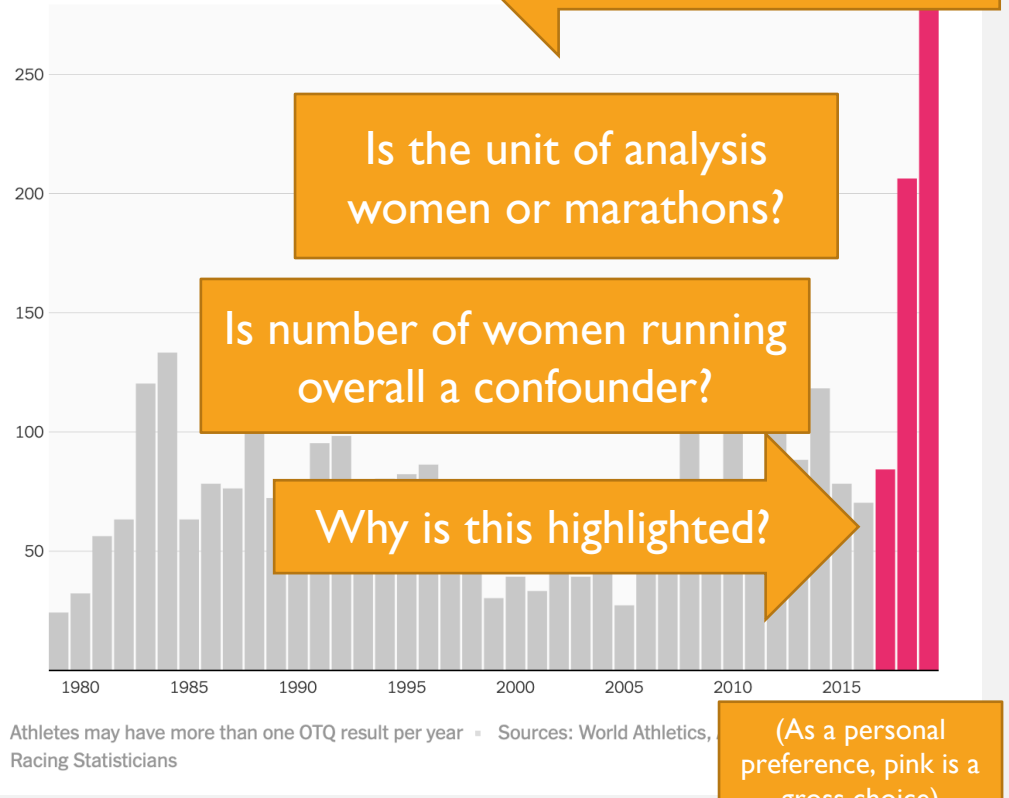
The women represent a range of backgrounds, and many have pushed themselves to results they once thought unachievable. We know this, because we talked to them — hundreds of them.

Another possible systematic bias in their responses?

Ok, but you can't claim anything causal

Through social media networks, running clubs, online message boards and the runners' own personal networks, The New York Times reached about two-thirds of the qualifiers.

Selection bias?!



Title could be clearer

Is the unit of analysis women or marathons?

Is number of women running overall a confounder?

Why is this highlighted?

(As a personal preference, pink is a gross choice)

IT'S NOT JUST THE NYT

Find the project here!

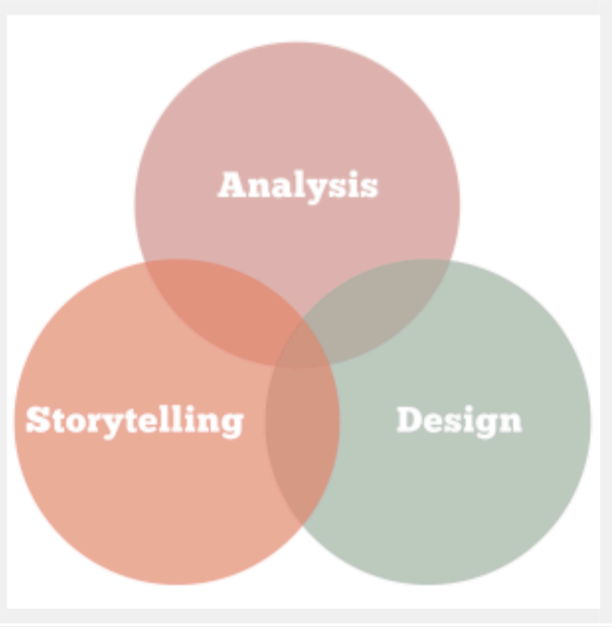
Outline

1. What is a data visualization?
2. Types of visualizations
3. What makes a good visualization?

Lecture 6.2:
An example from start through visualization
Functions & more helpful code

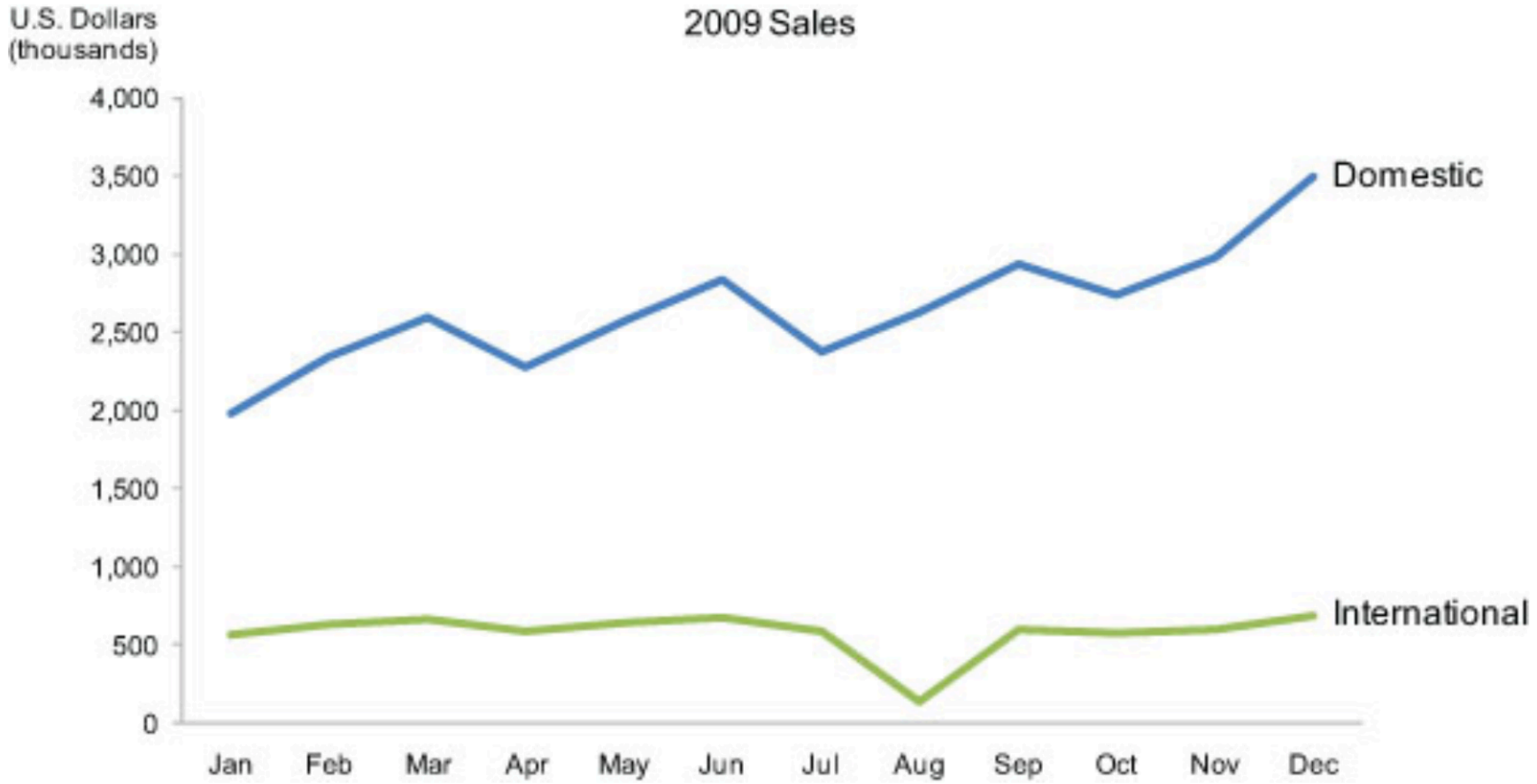
DATA VISUALIZATION

- The graphical display of abstract information
- For two purposes:
  - Sense-making
  - Communication
- Using principles of data analysis, design, and storytelling (yes)



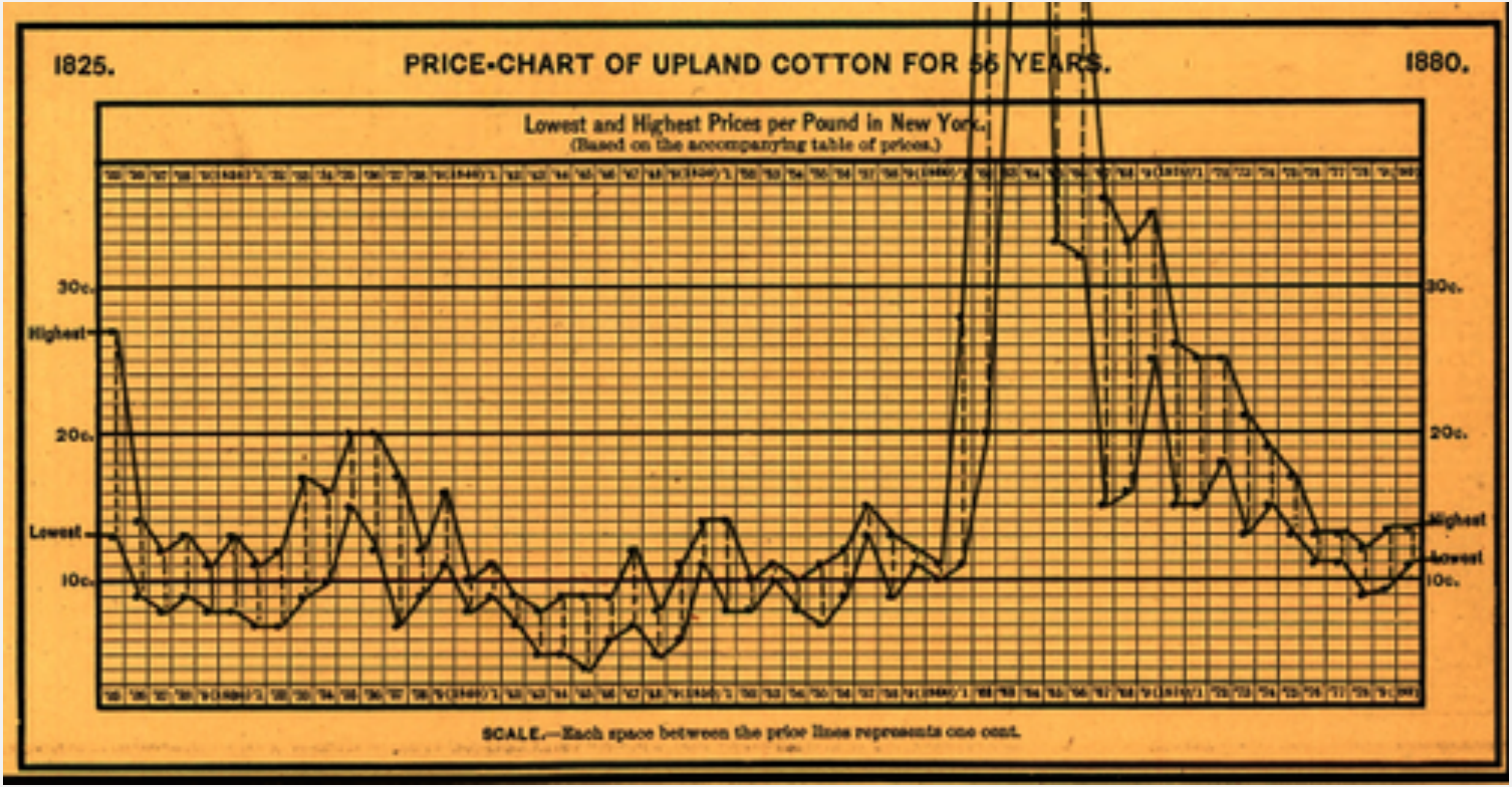
2009 Sales (thousands of U.S. $)

Region	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
Domestic	1,983	2,343	2,593	2,283	2,574	2,838	2,382	2,634	2,938	2,739	2,983	3,493	31,783
International	574	636	673	593	644	679	593	139	599	583	602	690	7,005
Total	2,557	2,979	3,266	2,876	3,218	3,517	2,975	2,773	3,537	3,322	3,585	4,183	38,788



DATA VISUALIZATION

- Again: Science and art
  - **Science**: How to measure, how to model, what are the interesting patterns or findings in this data and research?
  - **Art**: How to measure, how to model, what is interesting, and **how to capture or share what's interesting?**
- Something humans have been doing for a (relatively) long time!

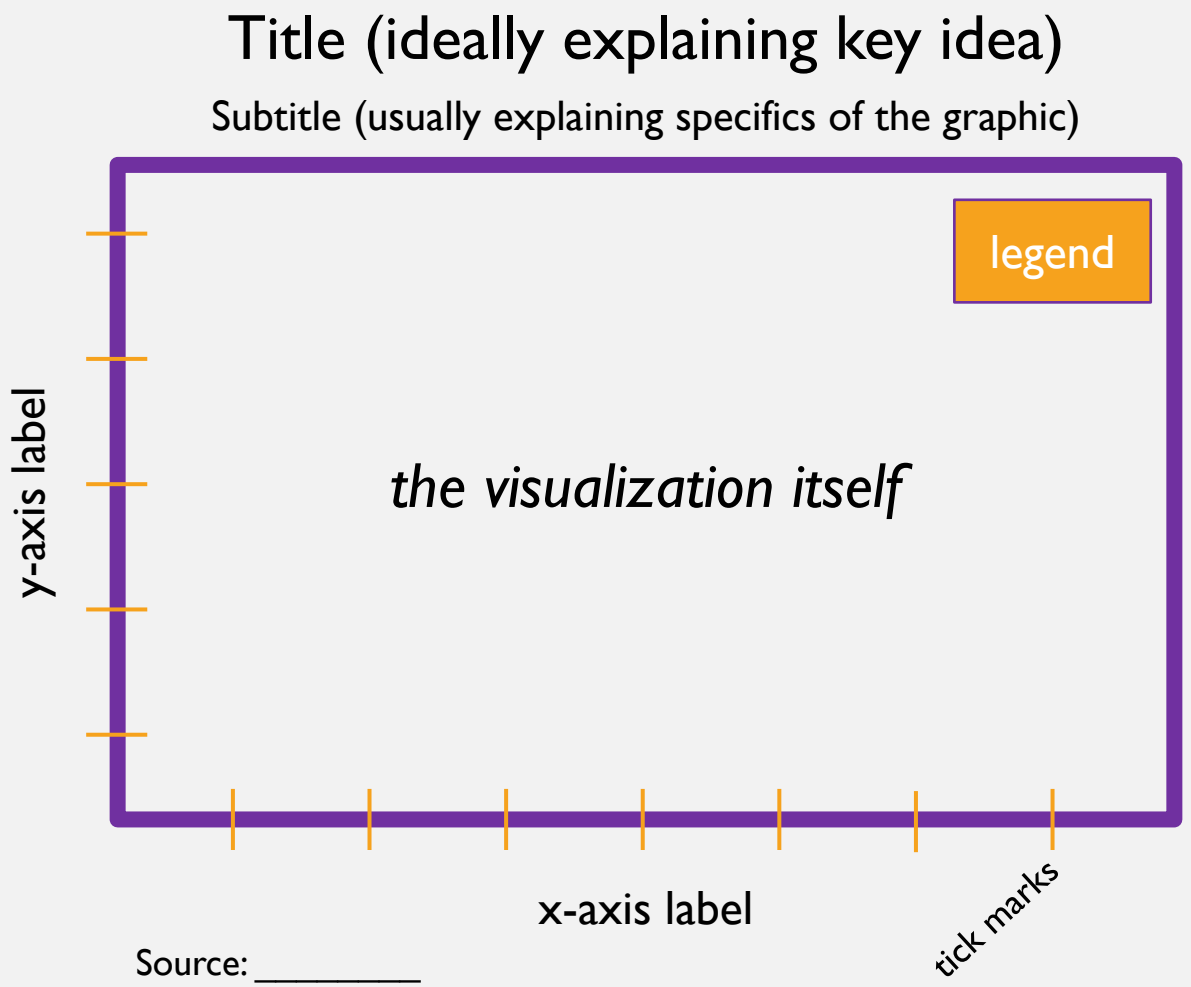


Source: handsomeatlas.tumblr.com

KEYS TO A GOOD VISUALIZATION

- Clear/understandable
- Ideally can stand on its own
- Well-labeled
- Draws attention to what's interesting/what the point is
- Honest: Doesn't mislead
- Transparent: about where the data comes from, how it's measured, what it really captures
- Replicable

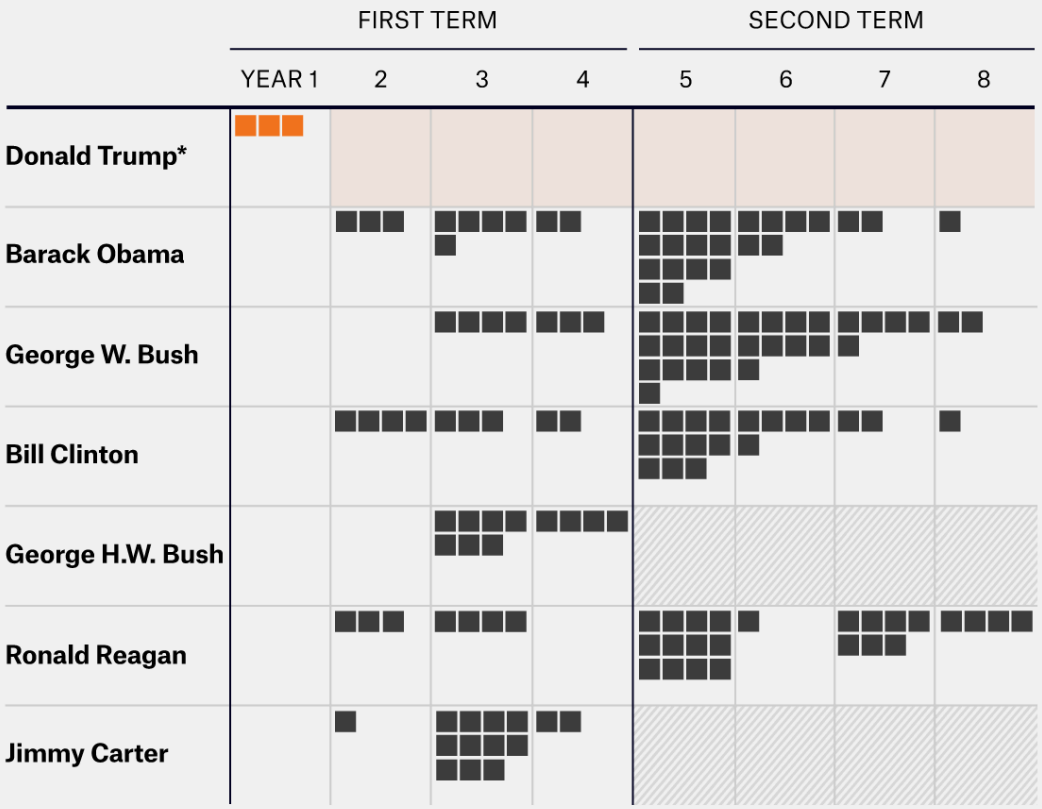
KEY ELEMENTS



Optional: Decoder of any additional information (e.g., colors)

Turnover among Trump's Cabinet has been high

Number of replacements in positions included in Trump's Cabinet, for recent presidents



*Trump's count is of departures because some replacements haven't happened yet.

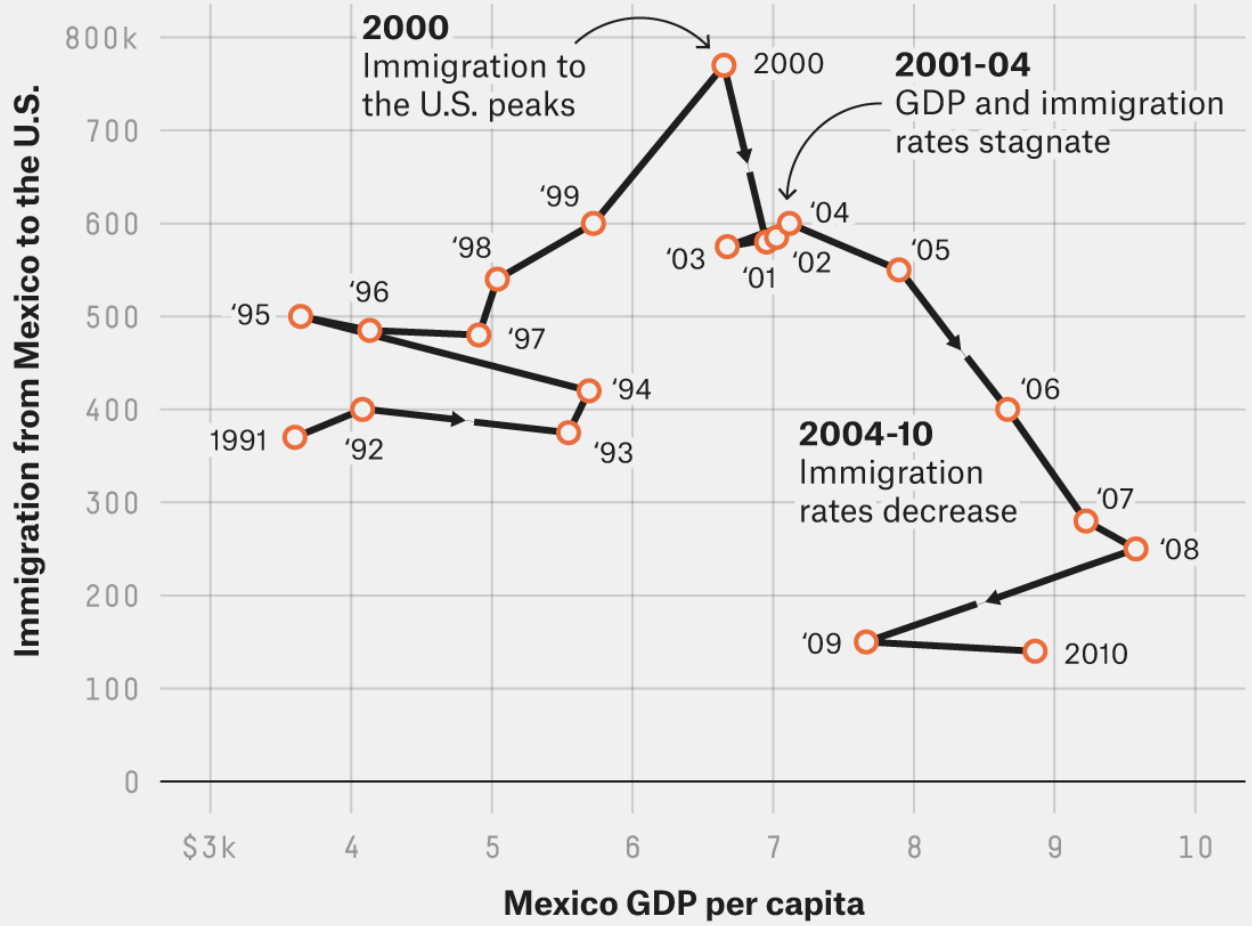
The number of total replaceable positions increases from 19 to 24 between Carter and Trump.

FiveThirtyEight

SOURCES: WHITE HOUSE, SENATE, DEPARTMENT WEBSITES

When the economy is up, immigration is down

Number of immigrants from Mexico to the United States versus Mexico's GDP per capita, 1991-2010



FiveThirtyEight

SOURCE: PEW RESEARCH CENTER, PEW HISPANIC CENTER, THE WORLD BANK

Outline

1. What is a data visualization?
2. Types of visualizations
3. What makes a good visualization?

FOUNDATIONAL VISUALIZATIONS

1. Bar charts
2. Line graphs
3. Scatter plots
4. Histograms
5. Others

Generating visualizations in Python may be simple or technically challenging

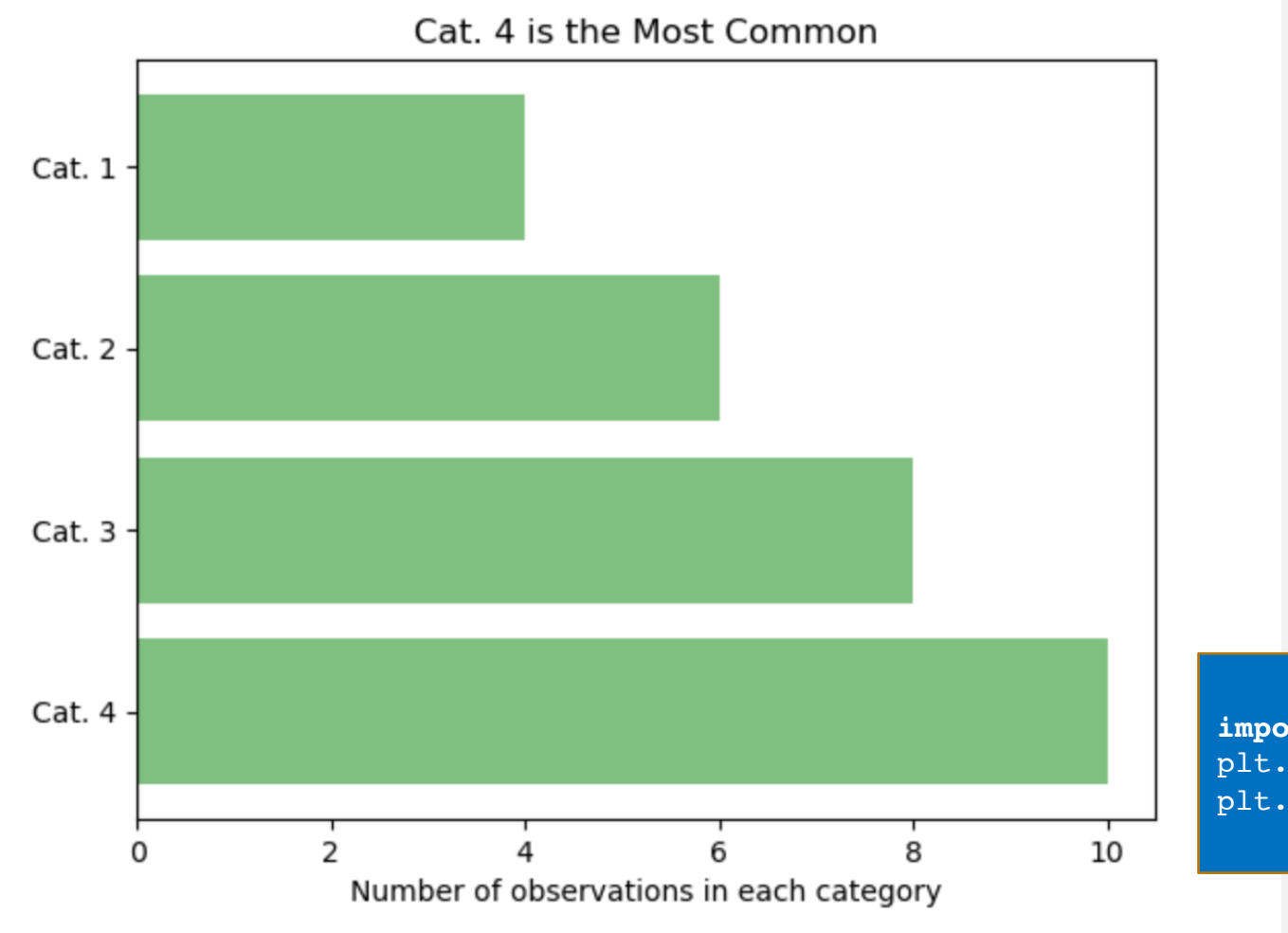
But a major challenge is often **determining the right visualization** in the first place

BAR CHARTS

Commonly used with **categorical variables**

Discrete, usually limited and finite, often **qualitative** property

Might also be useful for visualizing specific **observations**



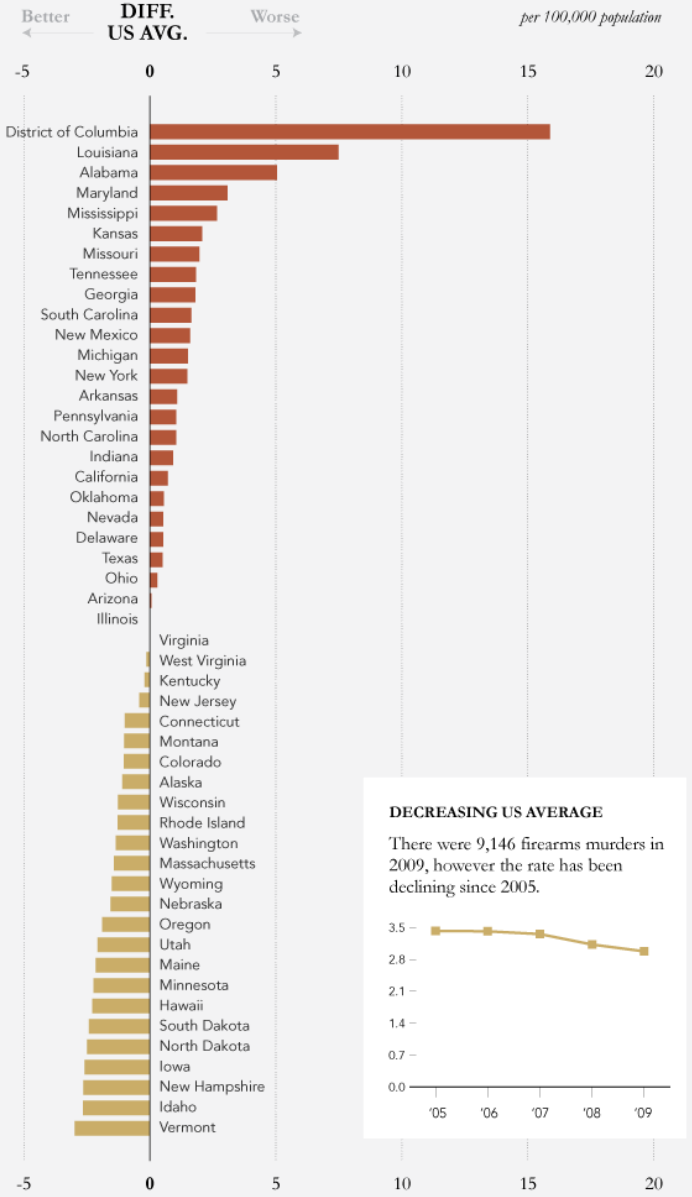
Examples

- Blood type: A, B, AB, O
- Hair color
- Type of book
- Political party
- Primary construction material
- Type of parole requirements
- Type of movie or award

```
import matplotlib.pyplot as plt
plt.bar(...) or plt.barh(...)
plt.show()
```

MURDERS BY FIREARM

The average rate for the United States was 2.98 murders by firearm per 100,000 population. The below shows how states compared to the national average.



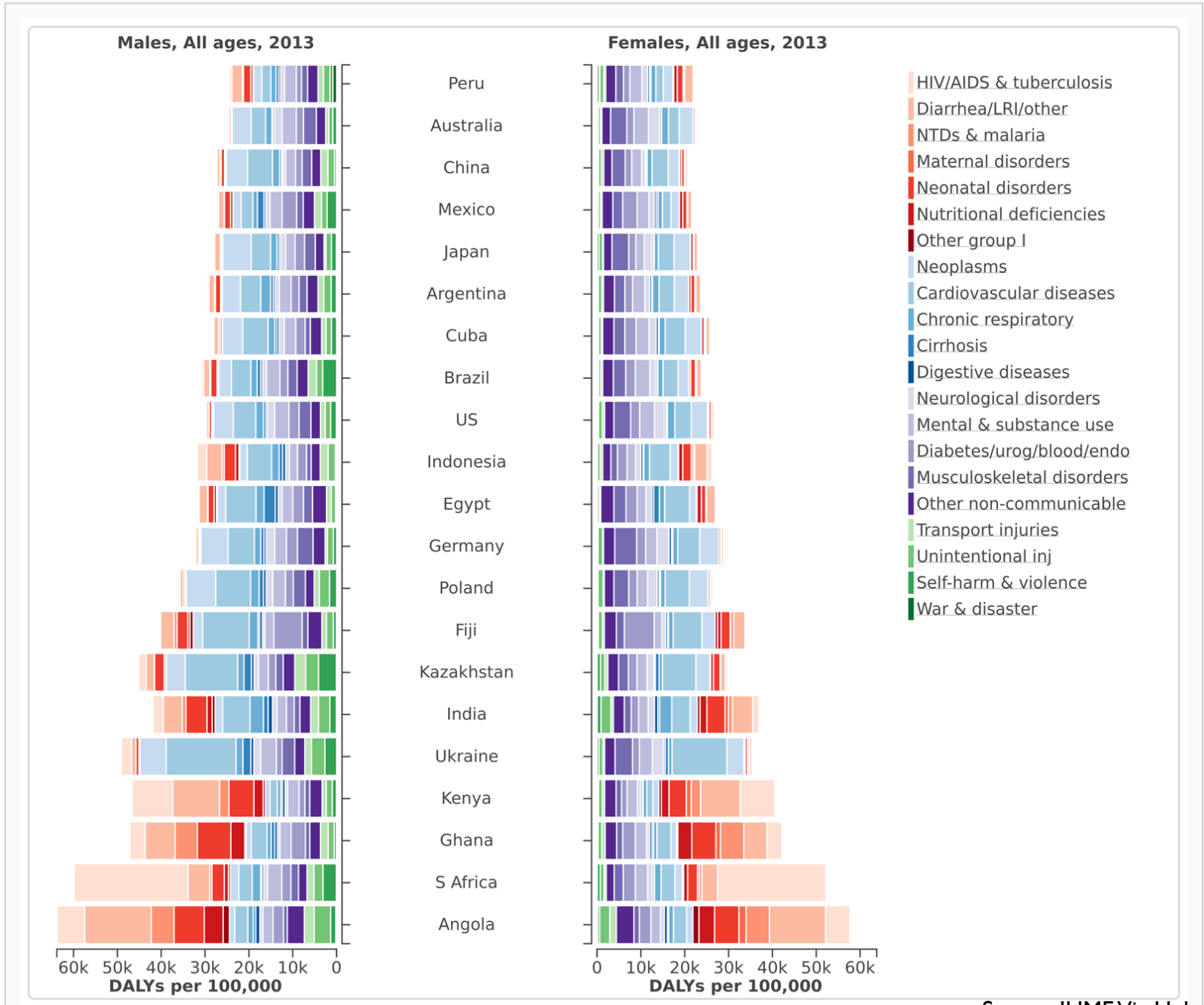
DECREASING US AVERAGE

There were 9,146 firearms murders in 2009, however the rate has been declining since 2005.

Source: Federal Bureau of Investigation | FlowingData, http://flowingdata.com

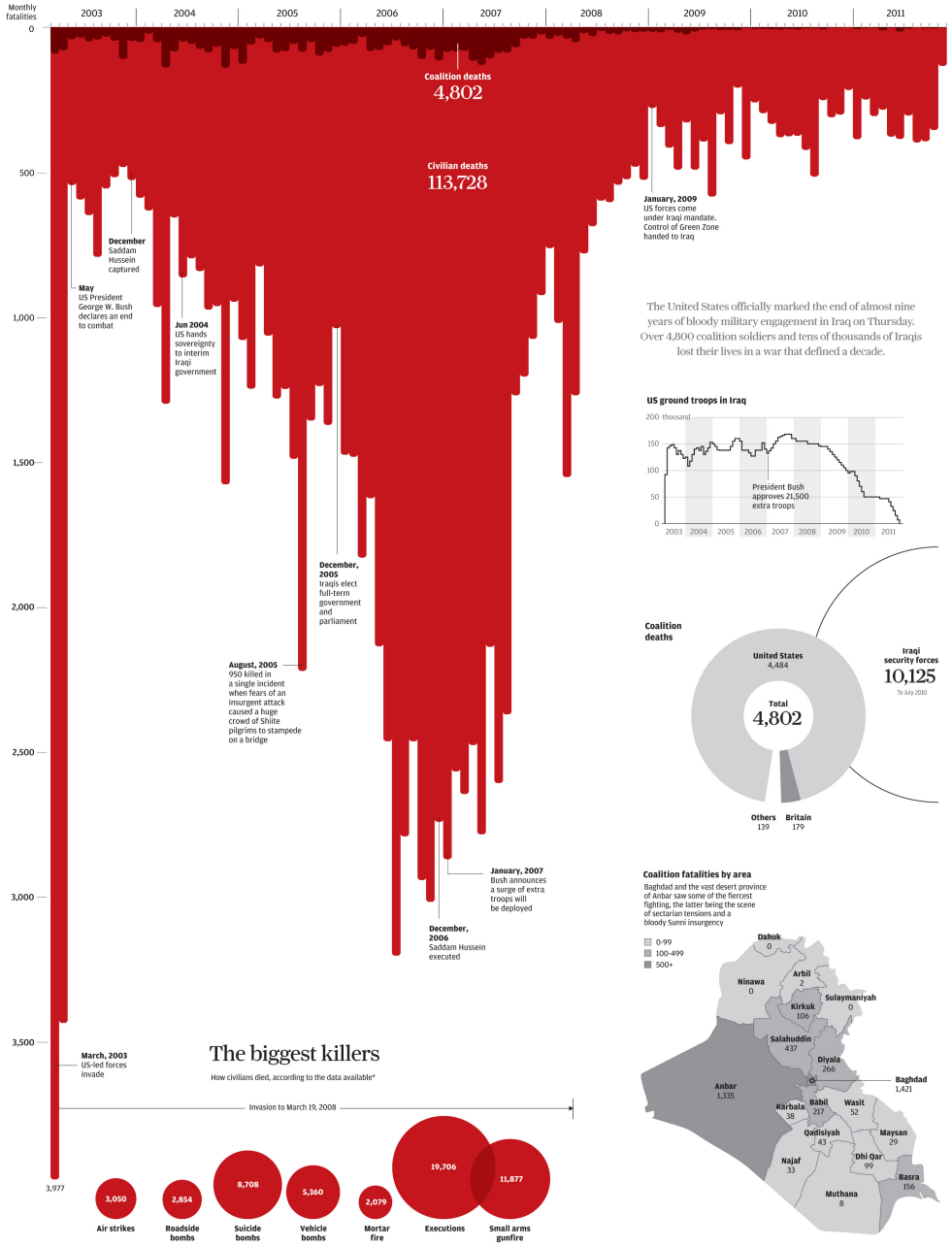
Source: FlowingData

Burden of disease by cause, country, and gender (2013 estimates) – produced by IHME Viz Hub



Source: IHME Viz Hub

Iraq's bloody toll



The United States officially marked the end of almost nine years of bloody military engagement in Iraq on Thursday. Over 4,800 coalition soldiers and tens of thousands of Iraqis lost their lives in a war that defined a decade.

SCMP graphic: Simon Scarr

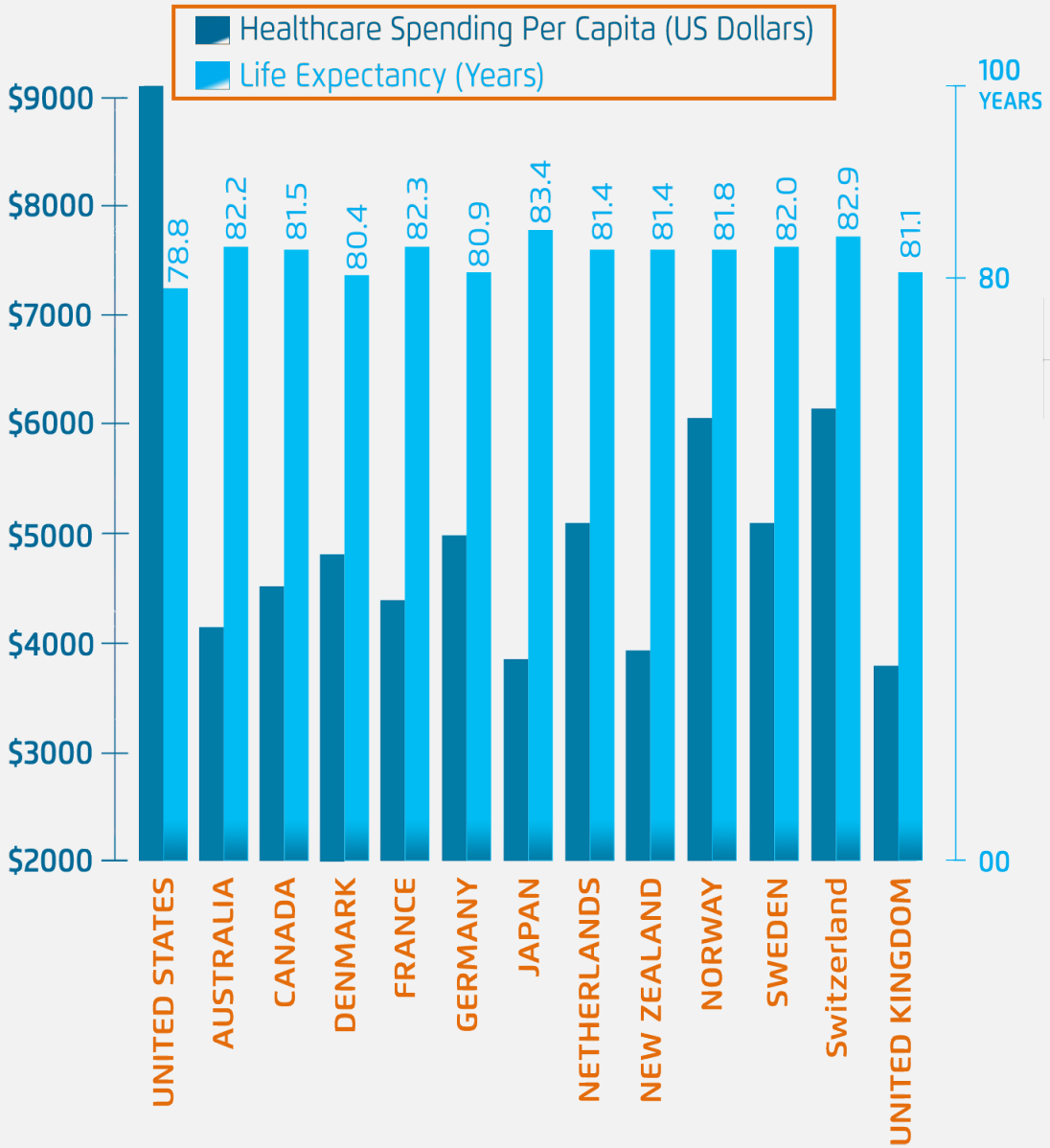
*Deaths from unknown causes are not included. Causes of death accounting for less than 0.5 per cent of killings also not included.

Sources: iCasualties, Iraq Body Count, New England Journal of Medicine, Global Security, Brookings Institution

Source: Washington Post

US Population Health

SPENDING ≠ POPULATION OUTCOMES



PERCENT GDP SPENT ON HEALTHCARE

United States	France	Germany	Switzerland	Canada	United Kingdom
17.1%	11.6%	11.2%	11.1%	10.7%	8.8%

UNITED STATES spends nearly 50% MORE than the next-highest country

PUBLIC $PENDING

US public spending
$4197 per Capita

UK public spending
$2802 per Capita

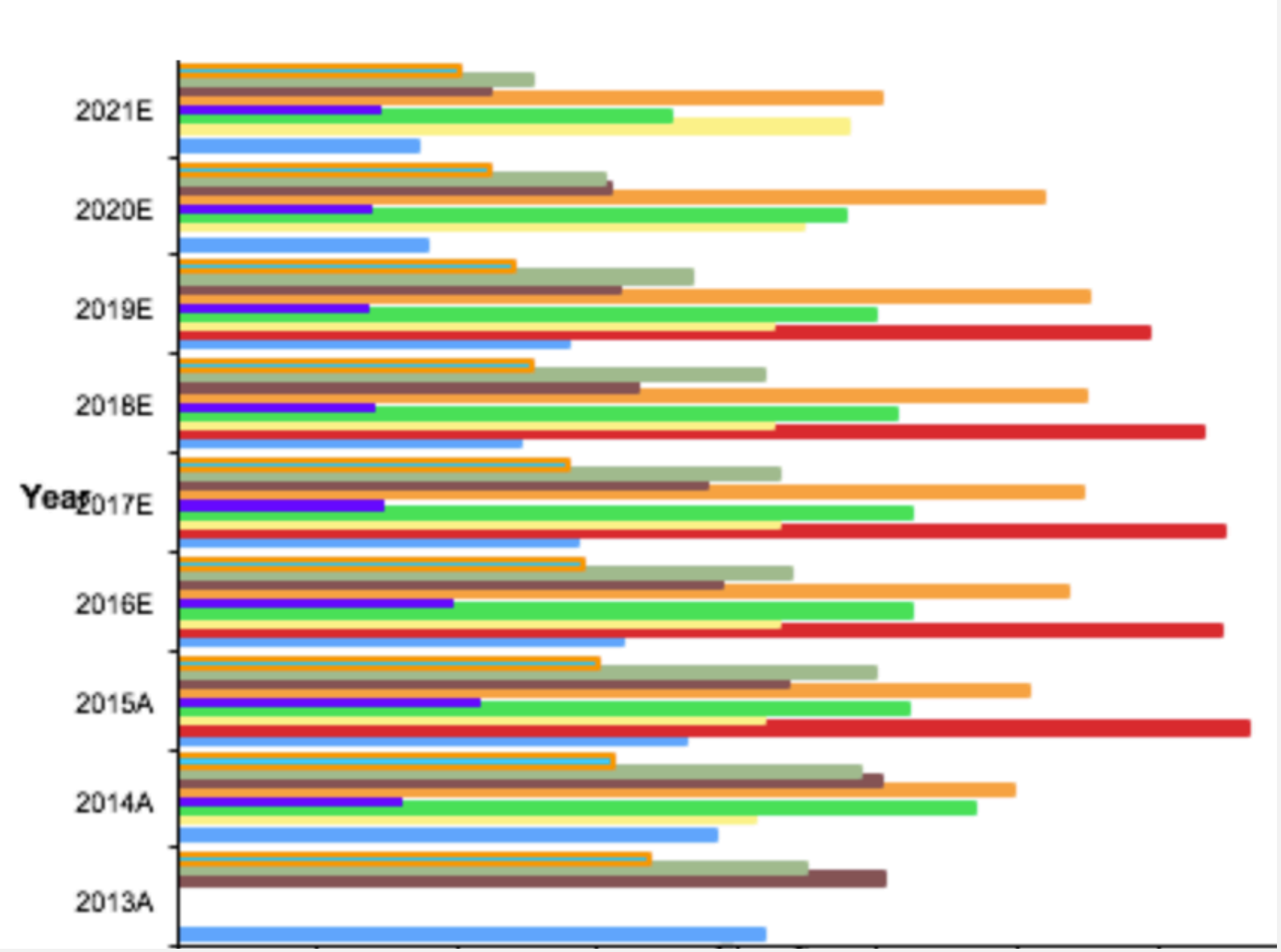
RESIDENTS COVERED BY PUBLIC SPENDING

US 34%

UK 100%

Source: Health Catalyst

Please don’t ever do this



Please explore these interactives!

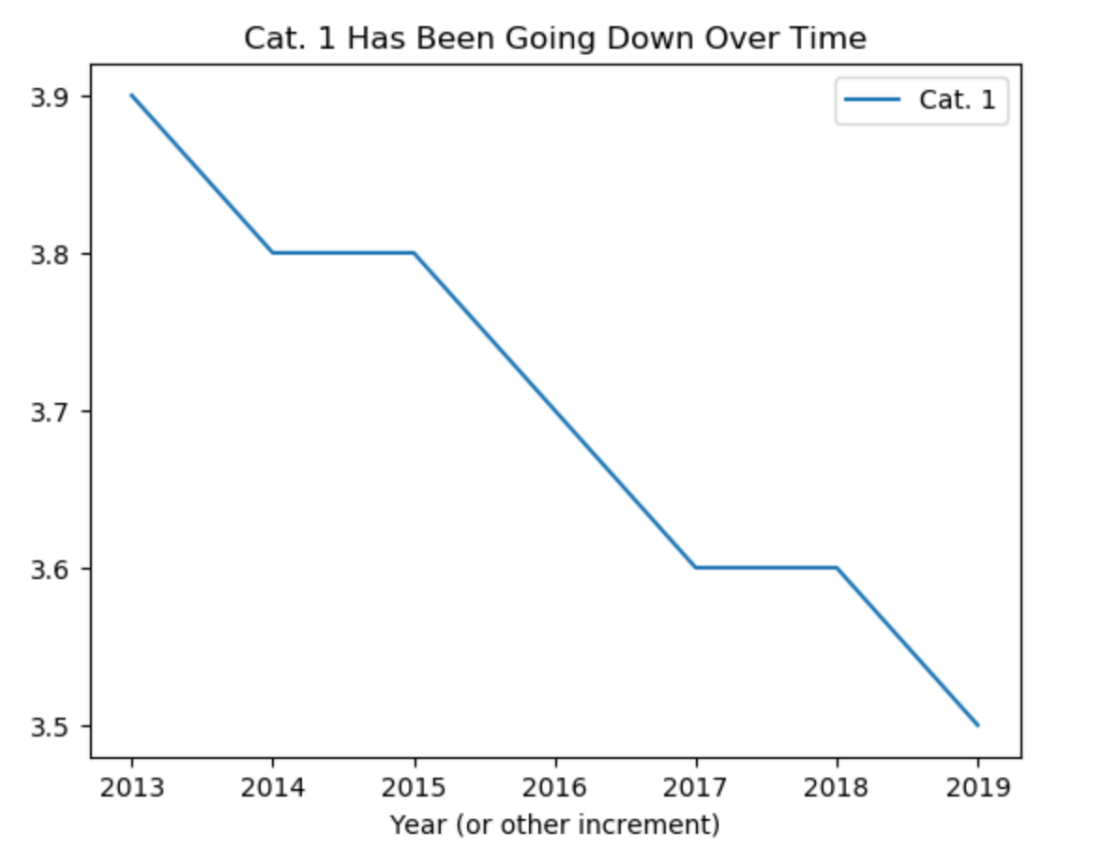
- Daily routines of famous creative people
- Distributions of annual incomes

LINE GRAPHS

Often used for chronological **trends** and patterns

Discrete or continuous, doesn't have to be over time, but is usually implied

Other continuous trends – temperature, speed, pressure readings



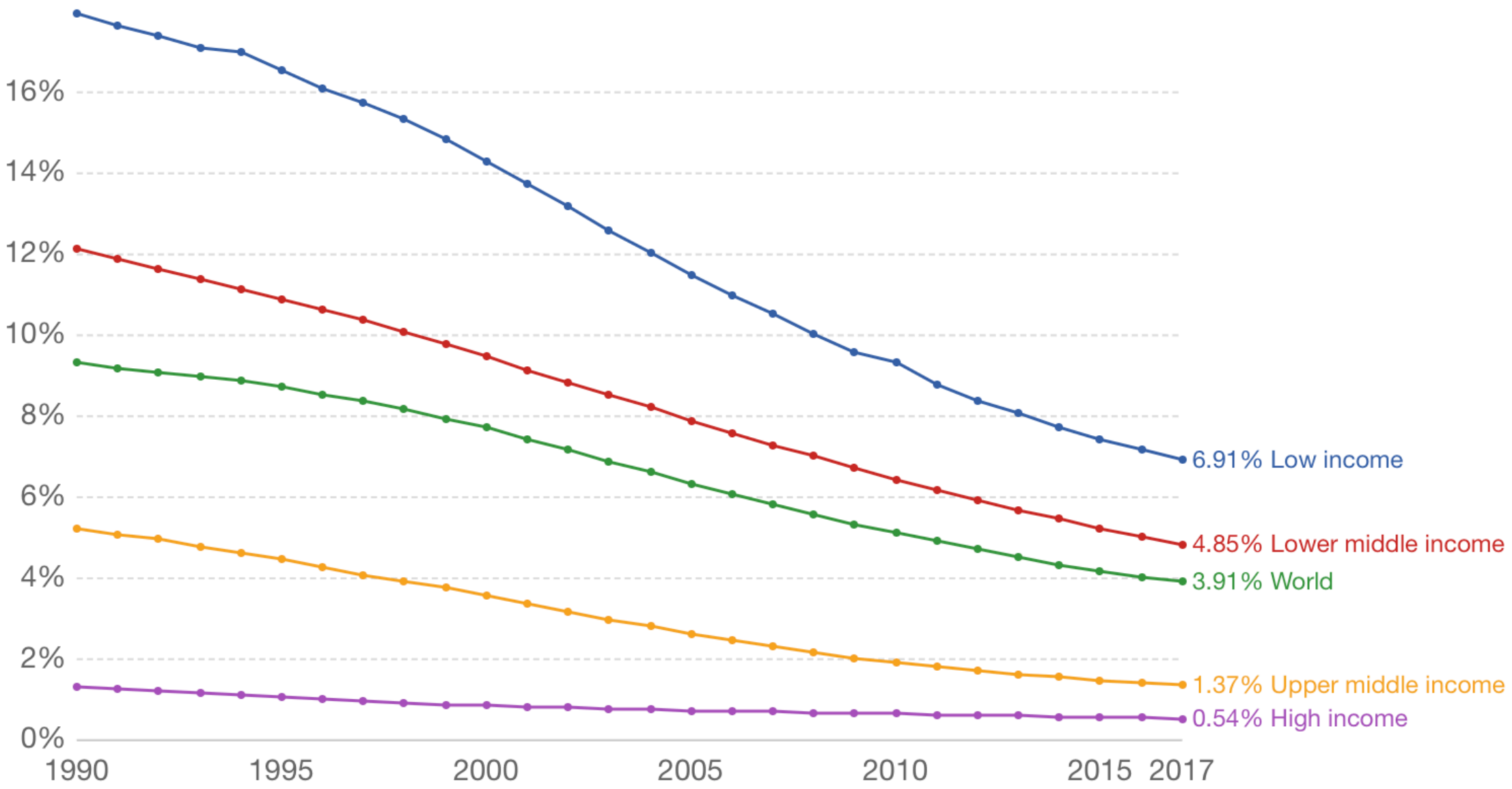
Examples

- GPA over time
- Voter turnout in elections historically
- Average rainfall per year in a location
- Effects of a change in temperature on a substance
- Pollution levels of a body of water over time

```
import matplotlib.pyplot as plt
plt.plot(...)
plt.show()
```

Child mortality by income level of country

The child mortality rate measures the share of children that die before reaching the age of 5.

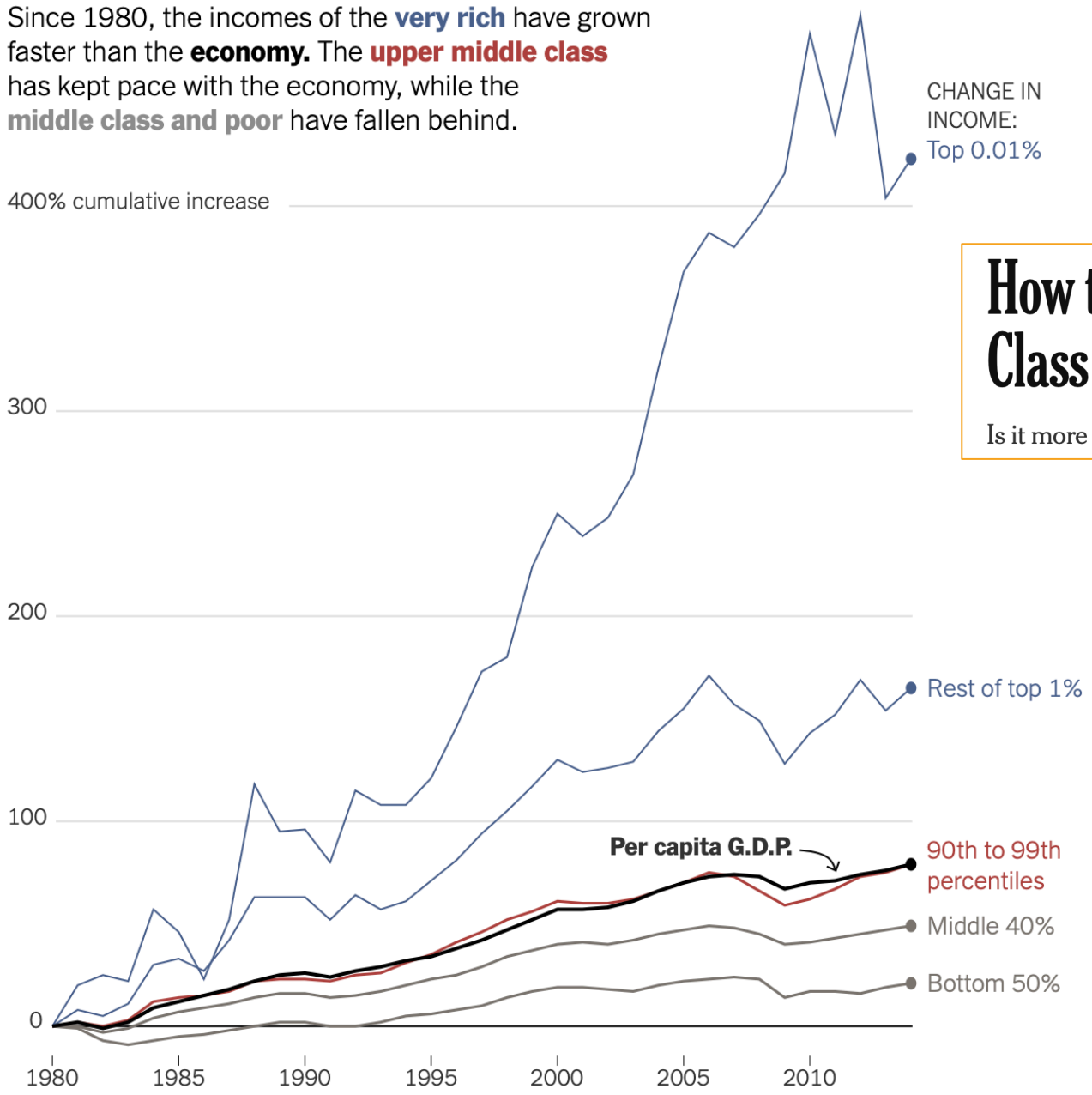


Source: World Bank

Source: Our World in Data

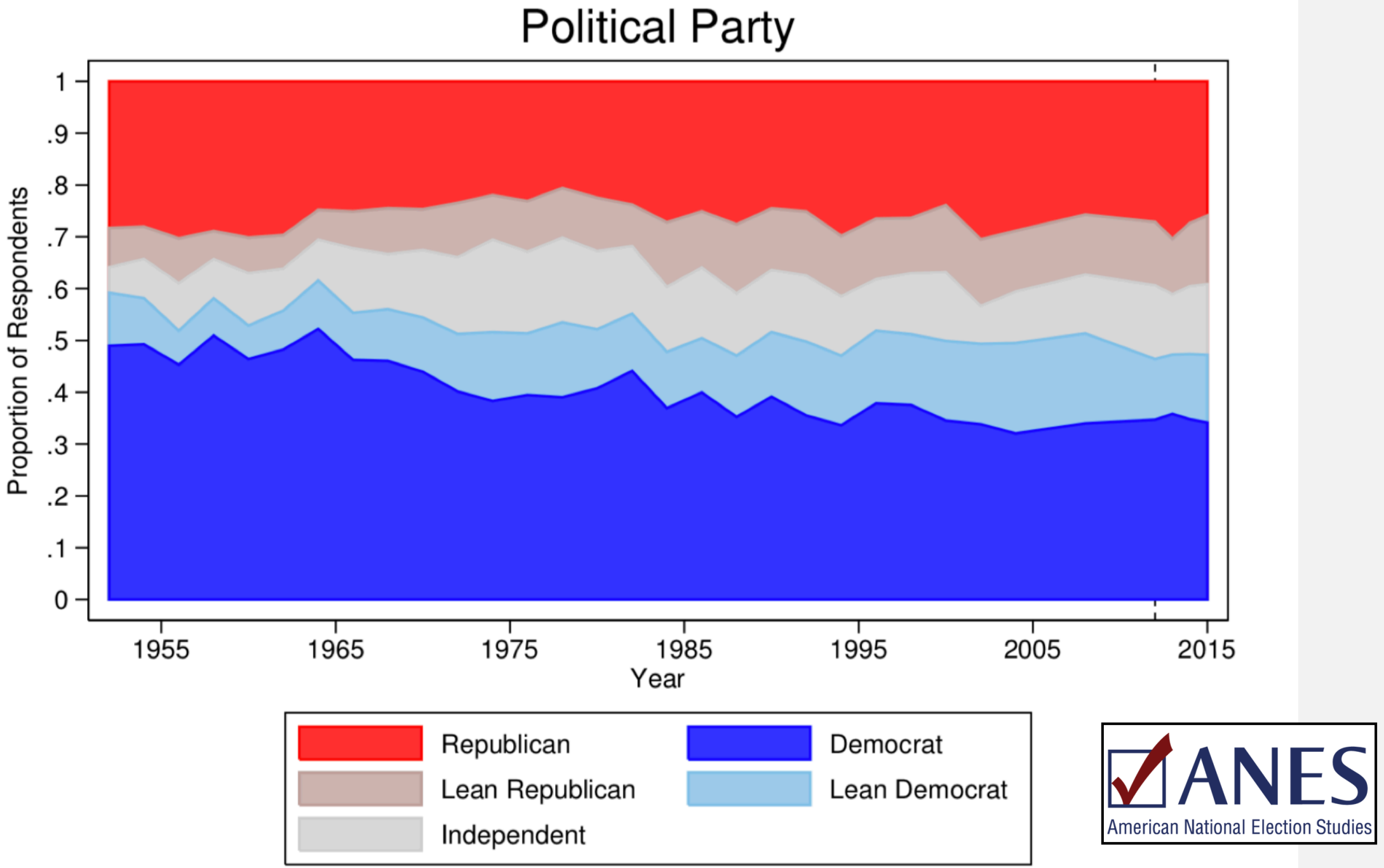
How the Upper Middle Class Is Really Doing

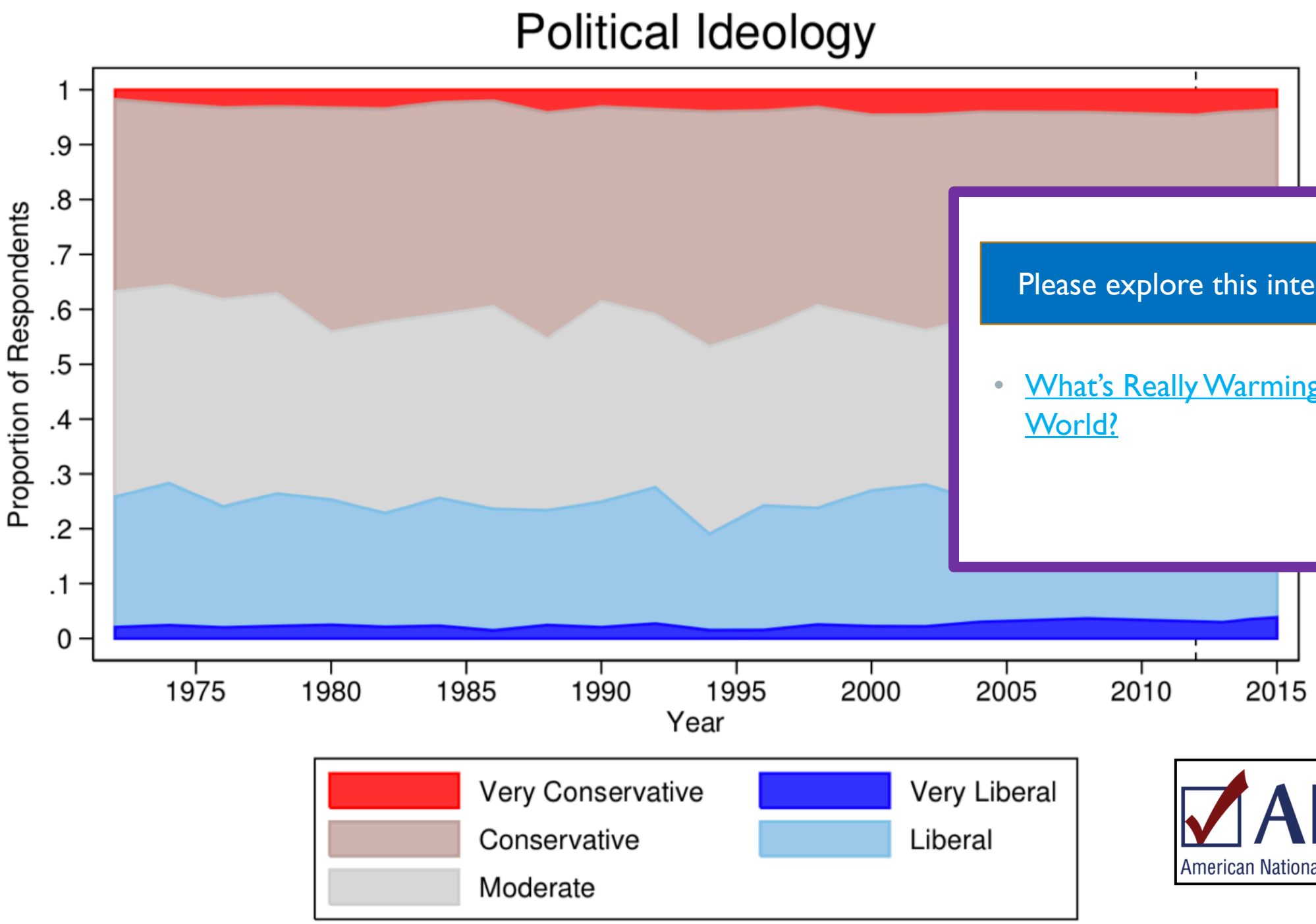
Is it more similar to the top 1 percent or the working class?



Note: Incomes are after taxes and include government transfers. • Sources: Thomas Piketty, Emmanuel Saez and Gabriel Zucman (incomes); Bureau of Economic Analysis (G.D.P.) • By The New York Times

Source: New York Times


Political Party
Proportion of Respondents
1
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
1955
1965
1975
1985
1995
2005
2015
Year
Republican
Lean Republican
Independent
Democrat
Lean Democrat
ANES
American National Election Studies



Please explore this interactive!

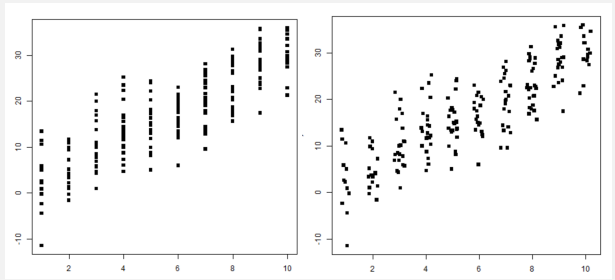
- What's Really Warming the World?

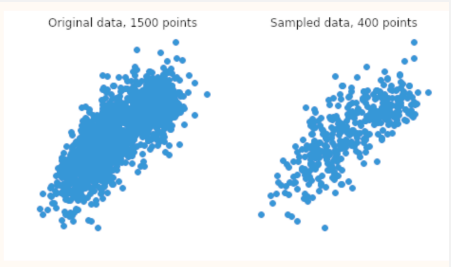
ANES
American National Election Studies

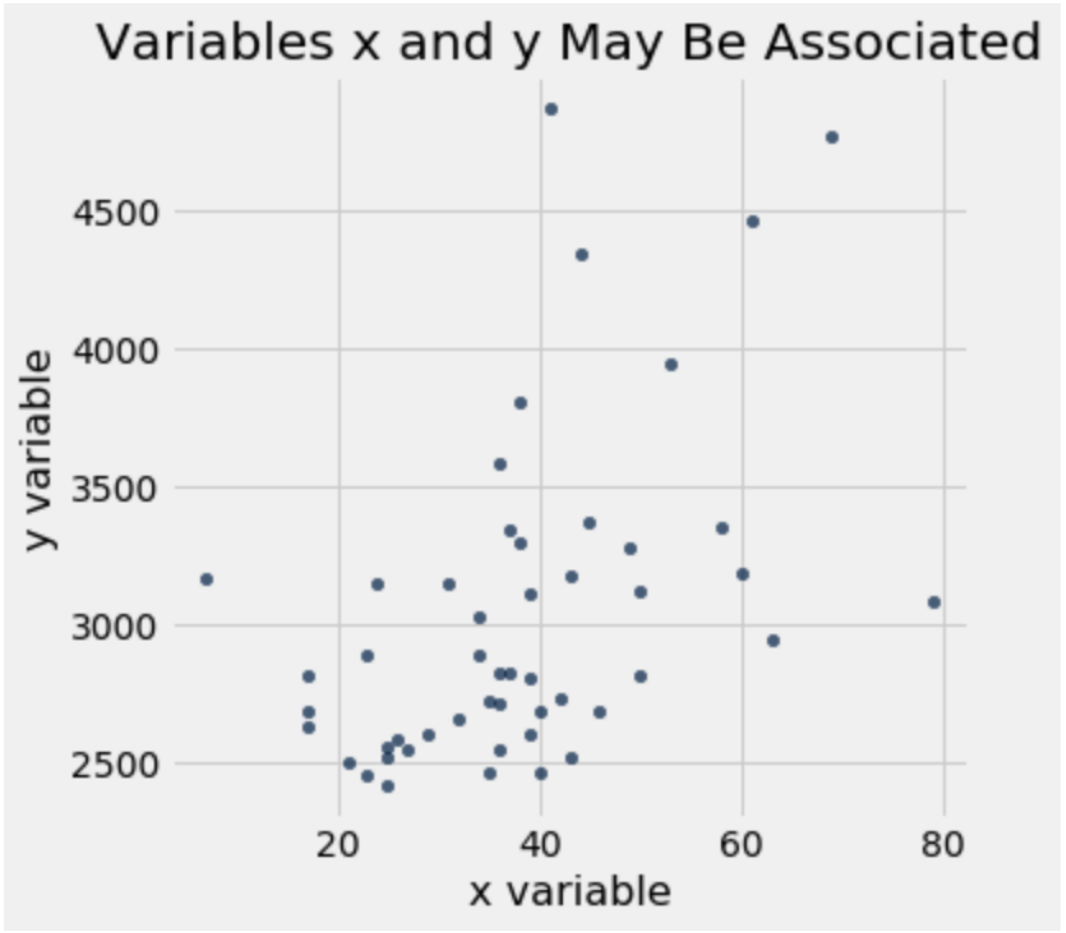
SCATTERPLOTS

Best for visualizing **association** between two variables

Continuous numeric variables tend to work best, though there are ways to work with others (e.g., jitter, select sub-sample)





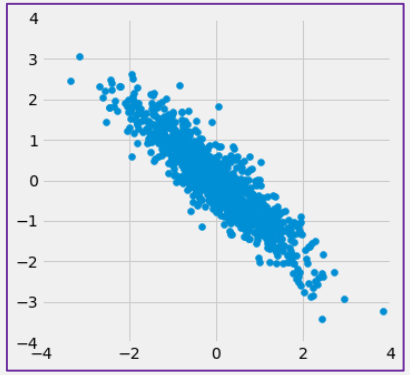
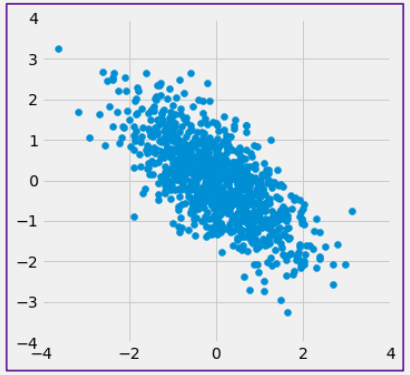
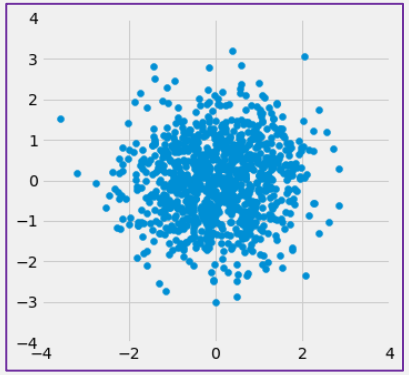
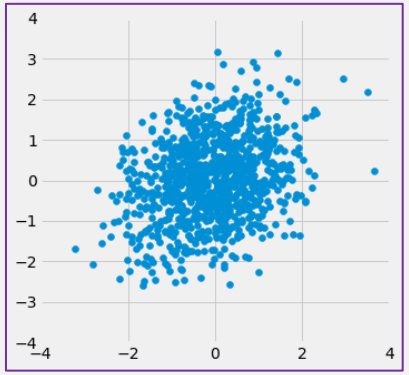
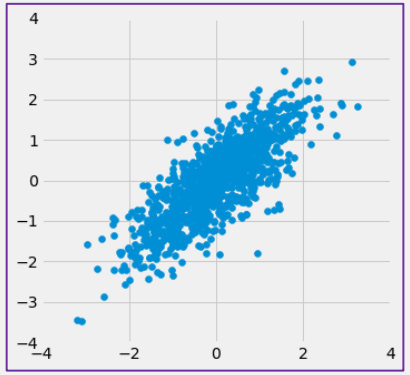
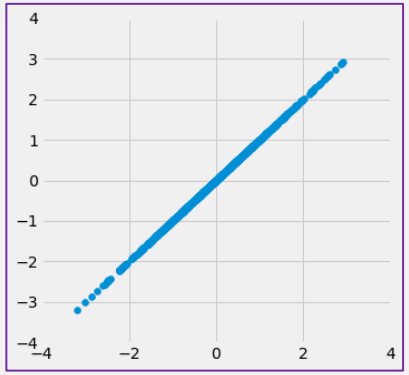


Examples

- Education and income
- Political stability and national GDP
- Coffee consumption and exam performance
- Social media usage and happiness
- Air pollution and hospital trips in a city

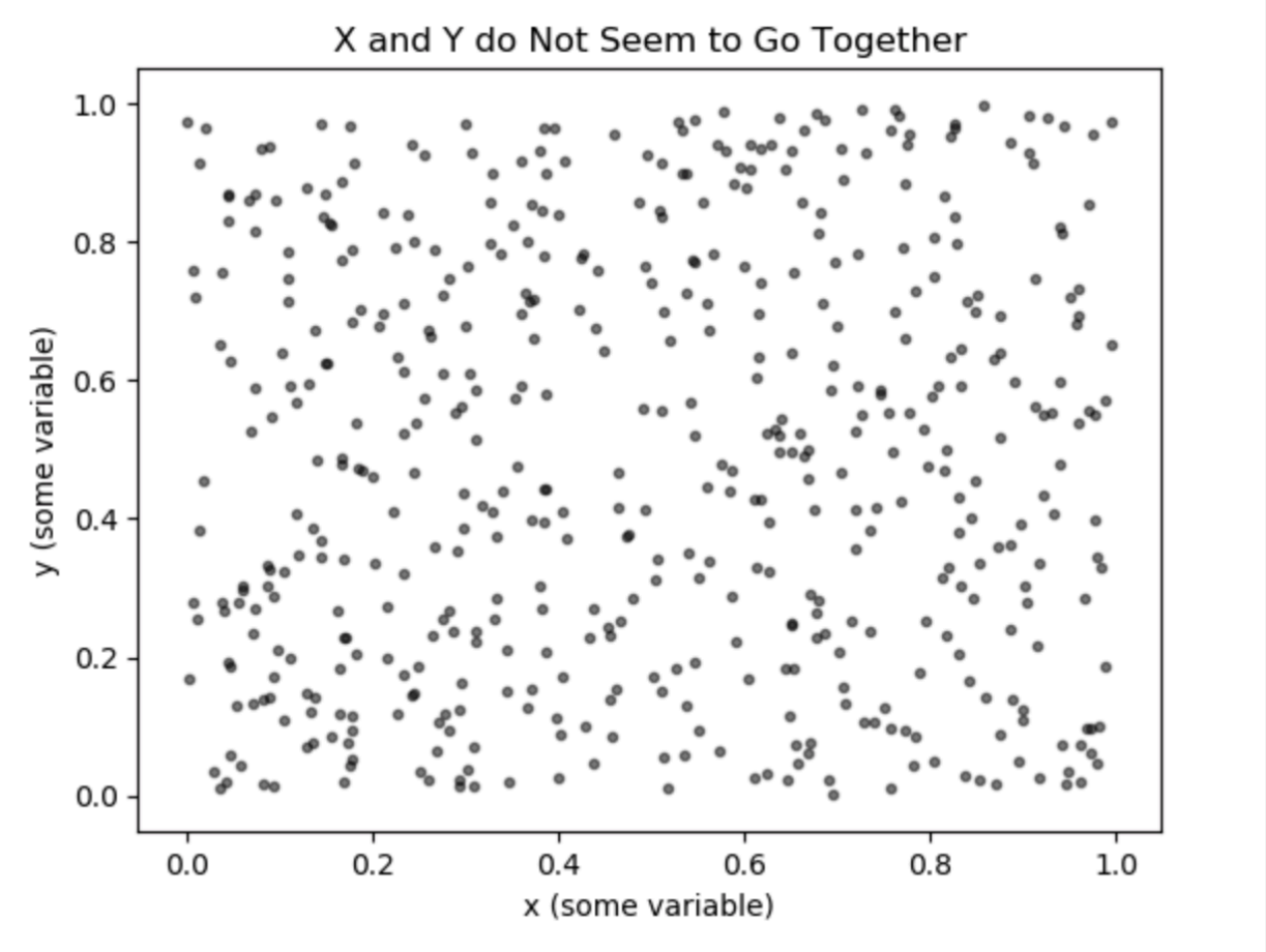
```
import matplotlib.pyplot as plt
plt.scatter(...)
plt.show()
```

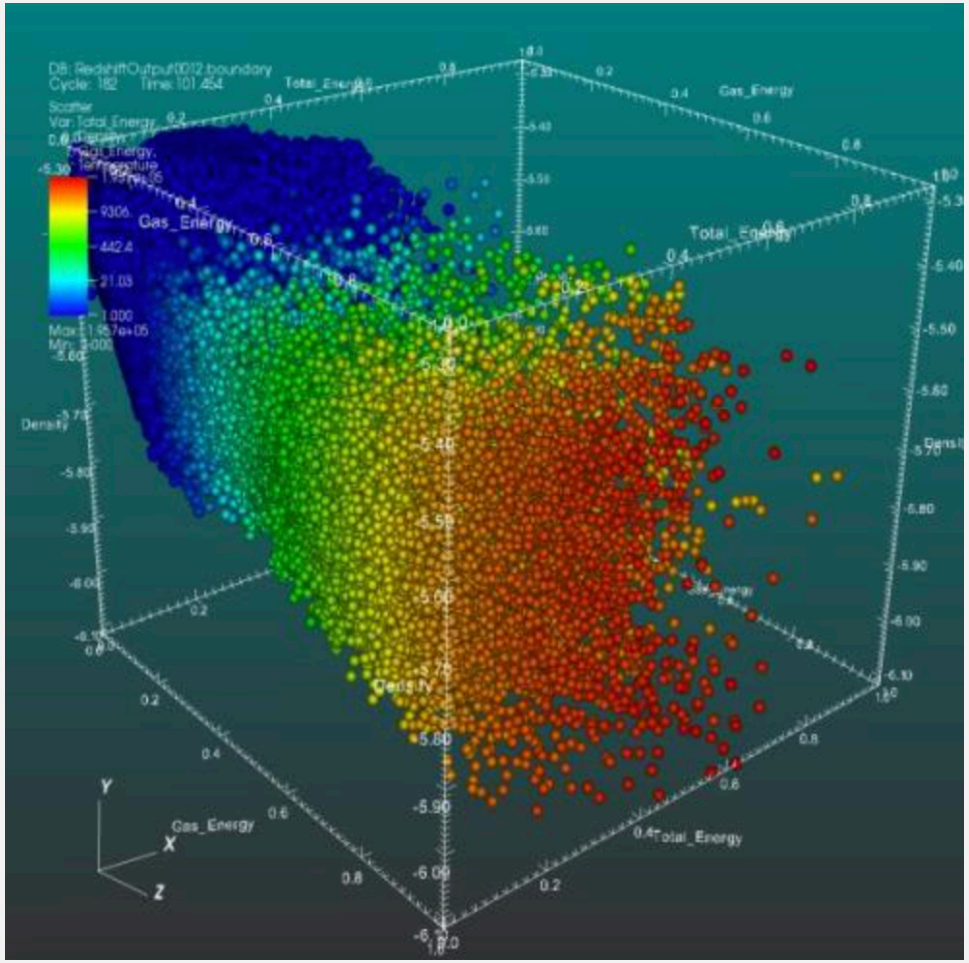
VISUAL INSPECTION OF ASSOCIATION BETWEEN TWO VARIABLES

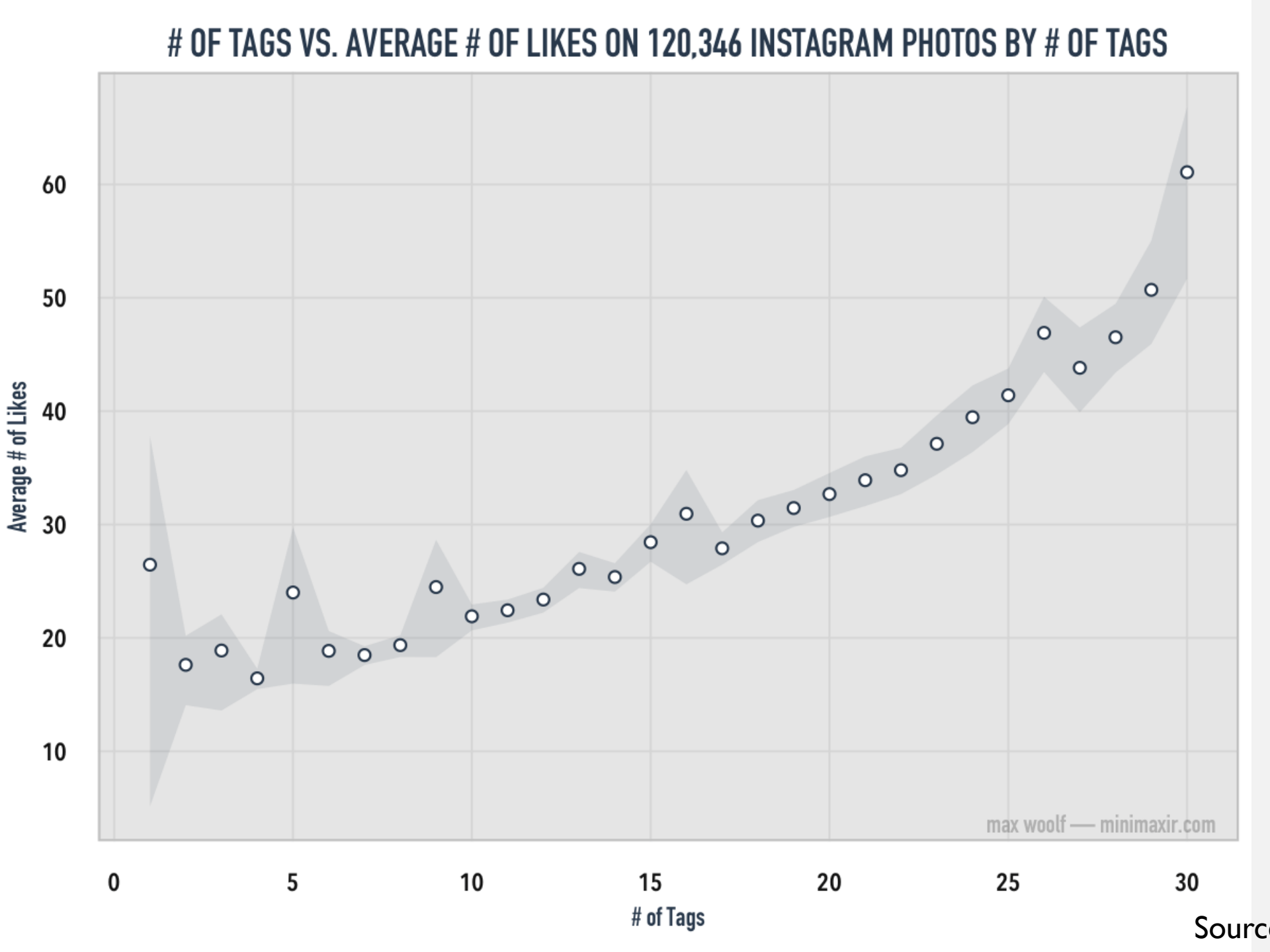







After the midterm, we will assign numbers to these relationships to quantify their correlation (*r*):

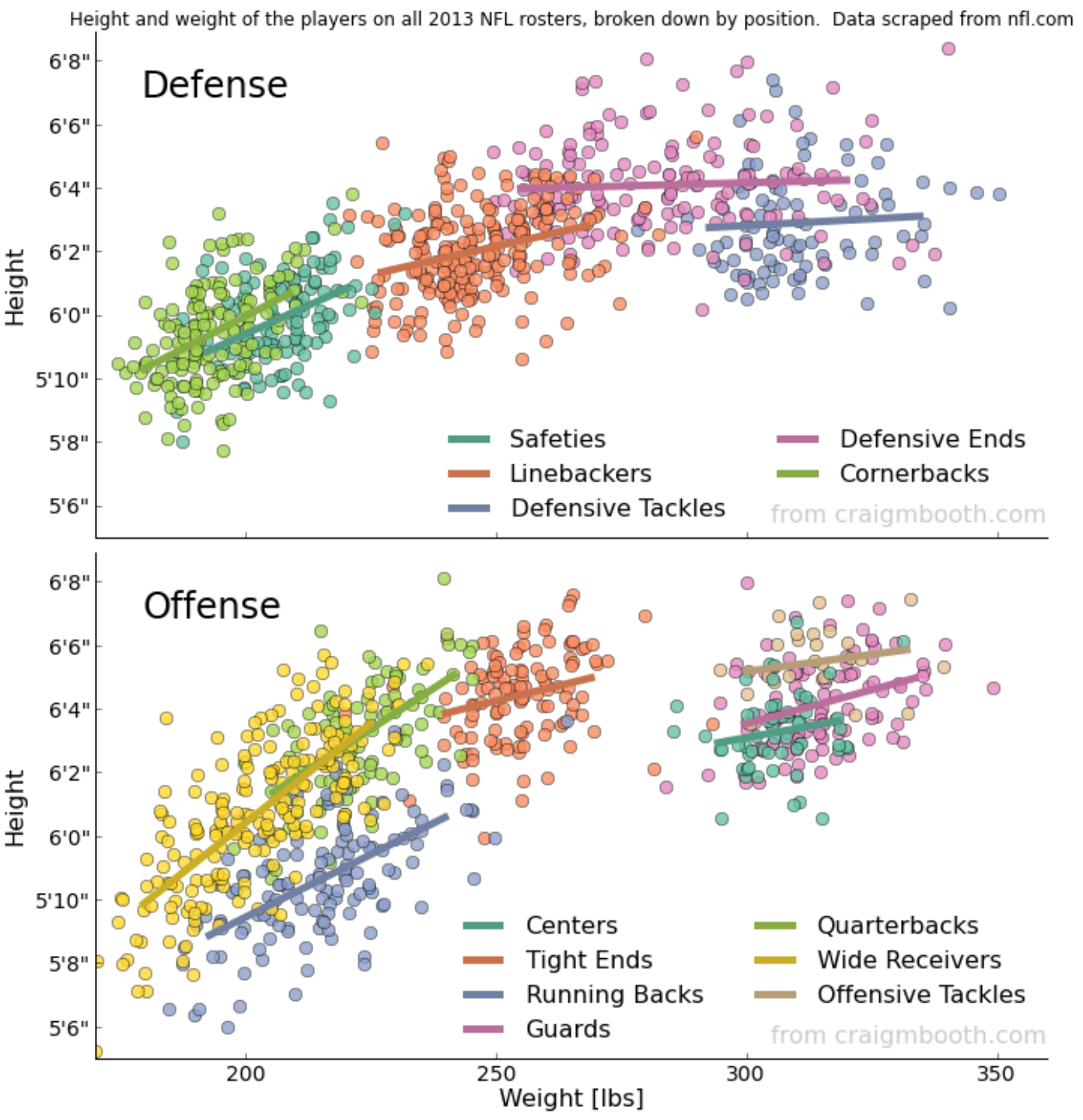
r= -0.9 r= -0.7 r= 0 r= 0.25 r= 0.8 r=1


X and Y do Not Seem to Go Together
y (some variable)
x (some variable)
0.0
0.2
0.4
0.6
0.8
1.0


Gas_Energy
Total_Energy
Density
X
Y
Z



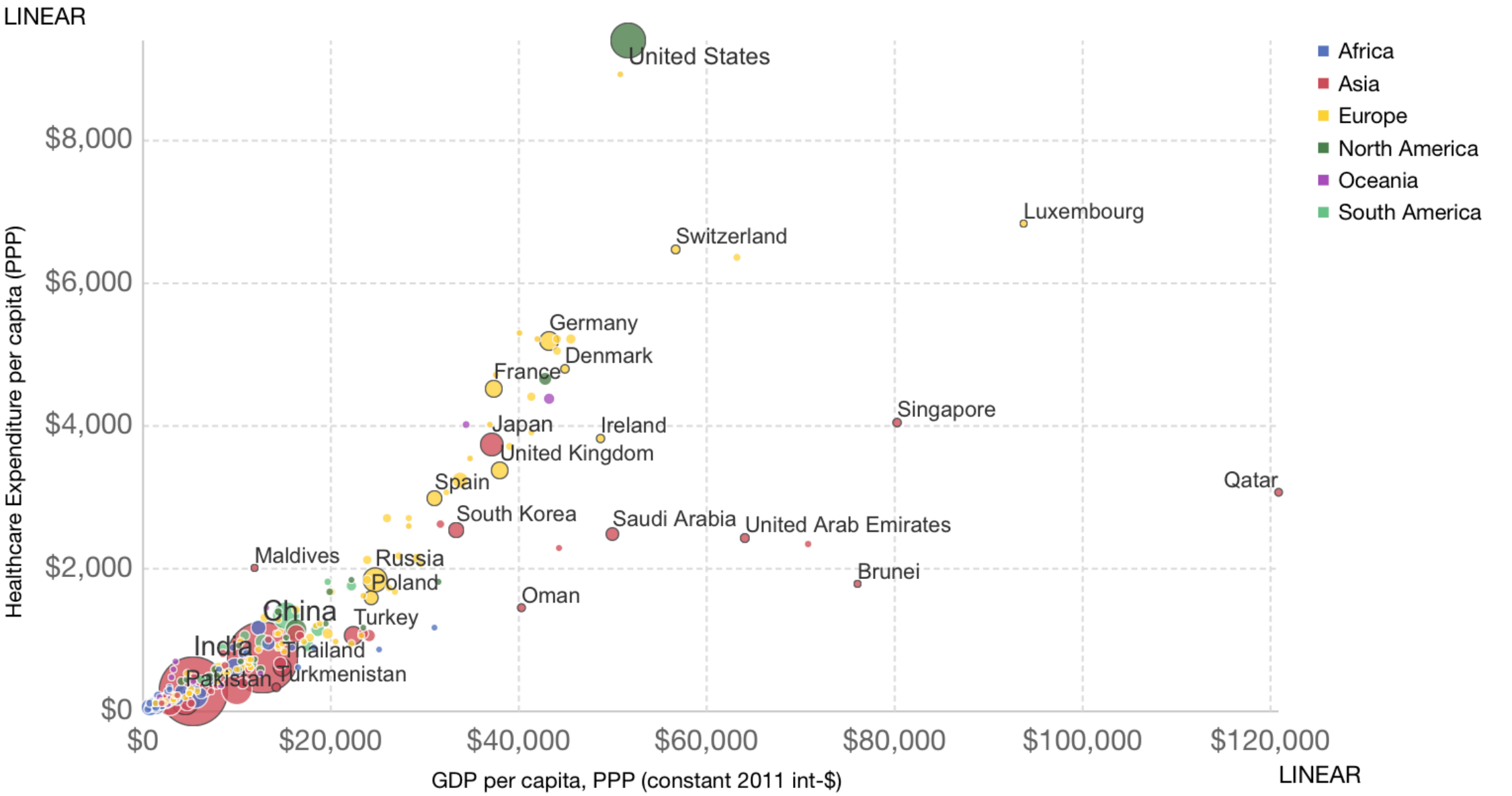
Source: Max Woolf



Source: Craig Booth

Healthcare Expenditure vs. GDP, 2014

Total healthcare expenditure per capita and GDP per capita in PPP-adjusted constant 2011 international dollar.



Source: World Bank – WDI

Source: Our World in Data

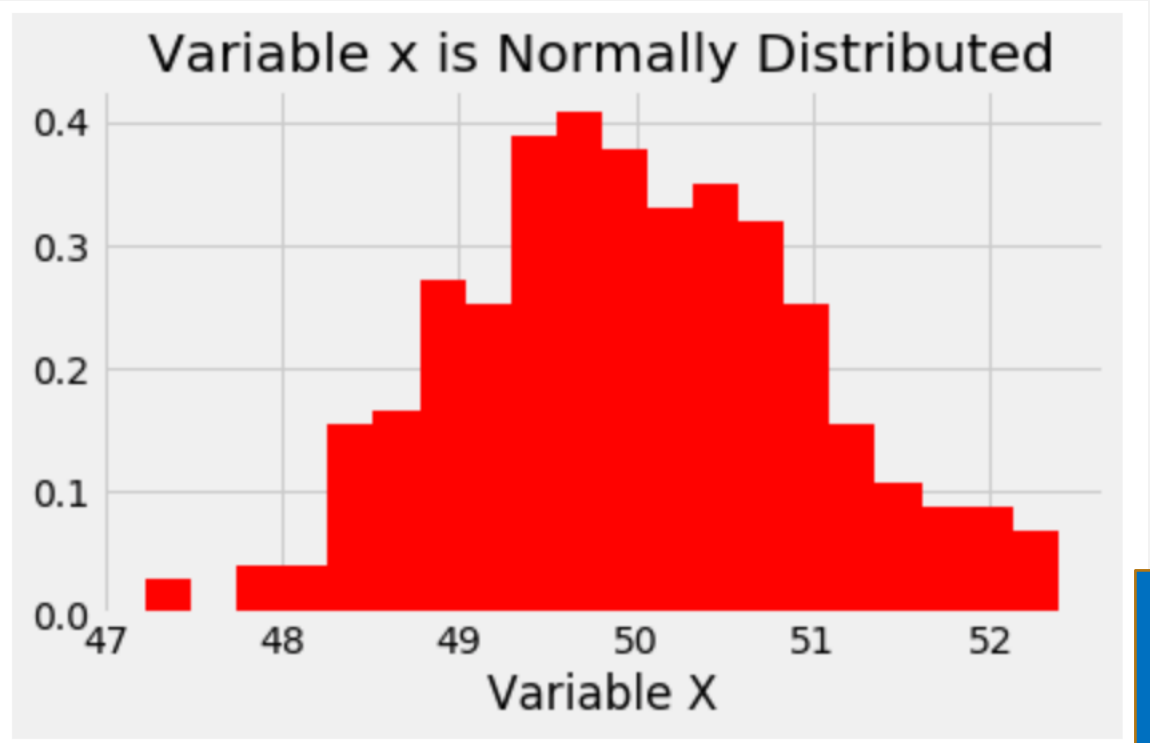
HISTOGRAMS

Used to understand the **distribution** of numerical variables (we already know this, too!)

Often for **continuous** data but discrete works, too

Data is broken up into groups called **bins**

The distribution of a variable is useful for understanding and predicting

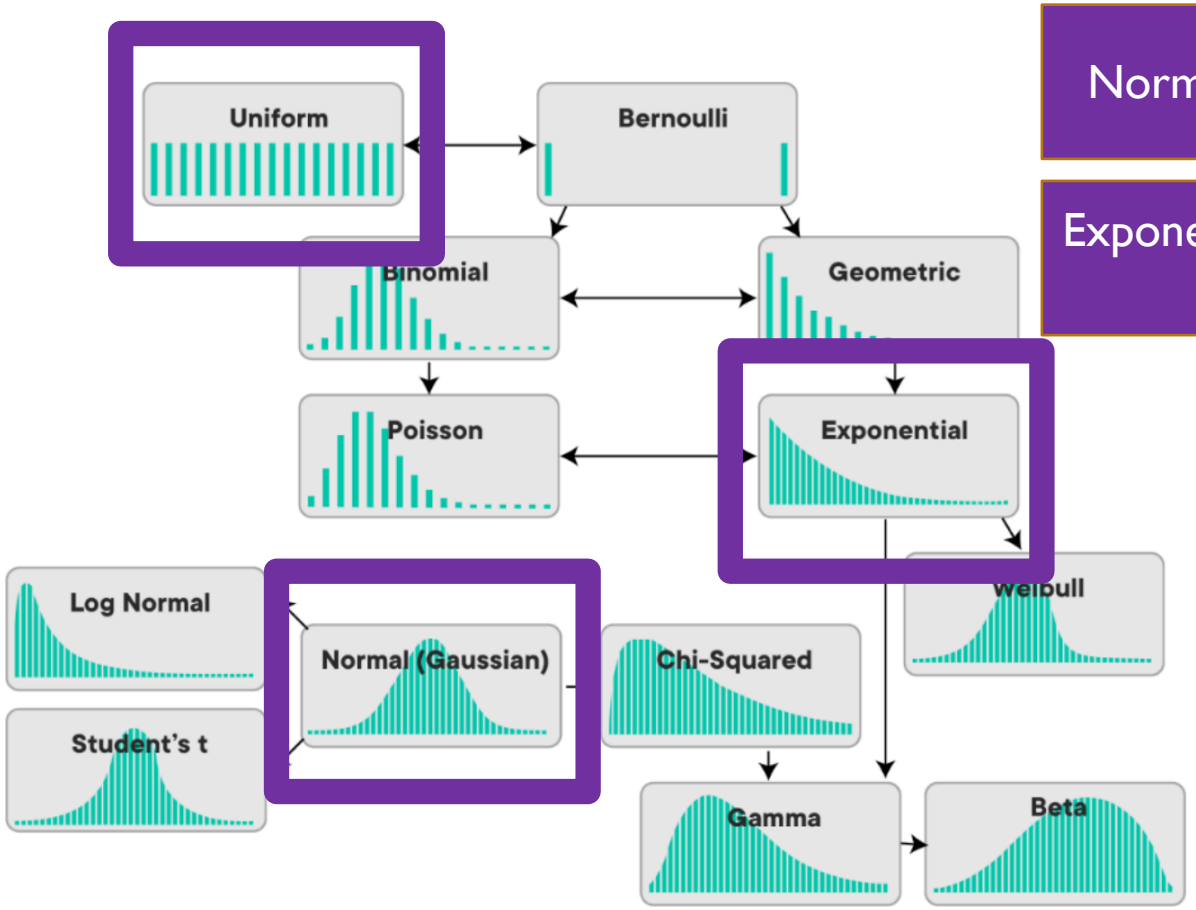


Examples

- Height
- Income
- Age
- How many copies of books sold
- Size of earthquakes

```
import matplotlib.pyplot as plt
plt.hist(...)
plt.show()
```

Understanding Different Types of Distributions You Will Encounter As A Data Scientist



Common types of distributions

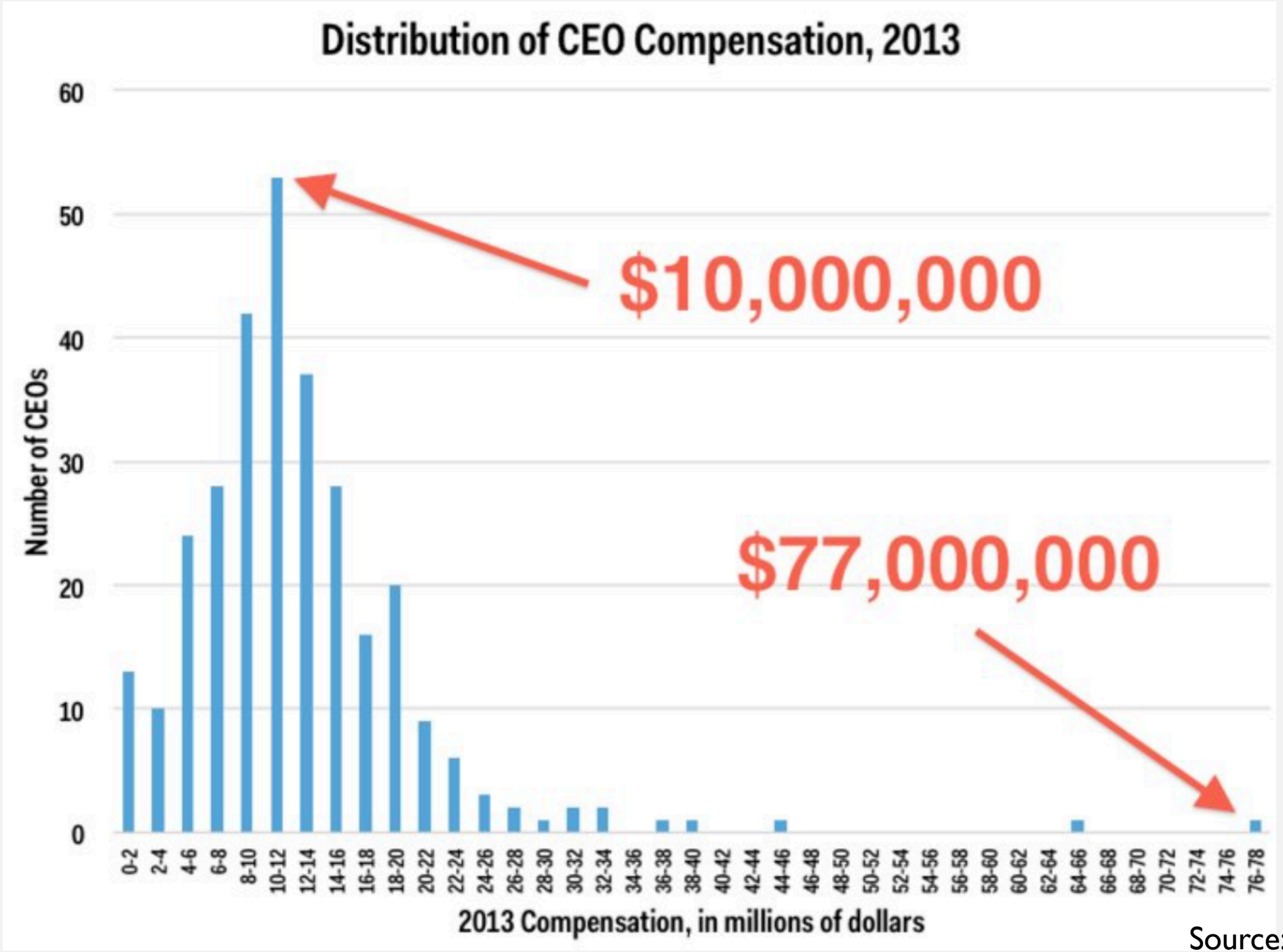
Uniform: All outcomes are equally likely

Normal: Middle values are most likely

Exponential: Low values most likely, high values very unlikely

Power law: Even more extreme

Read more here!



Source: Business Insider

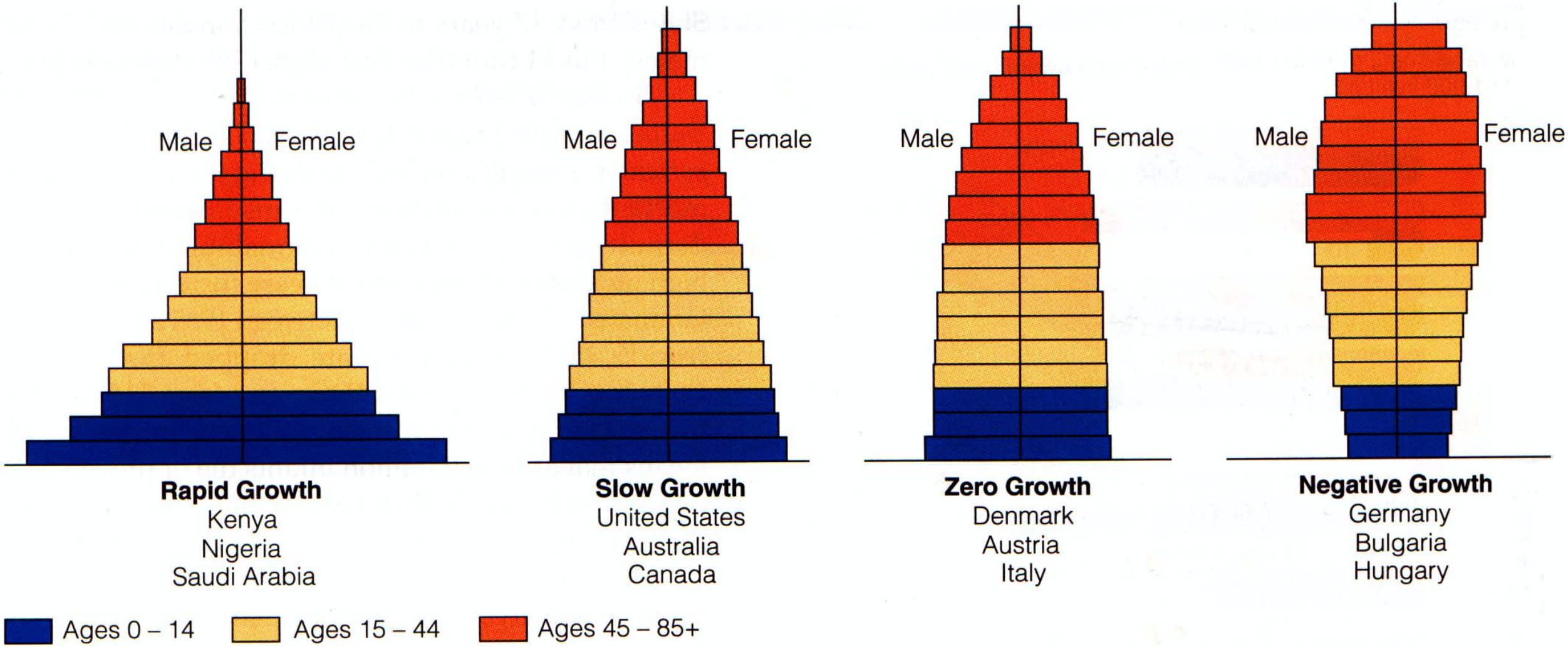
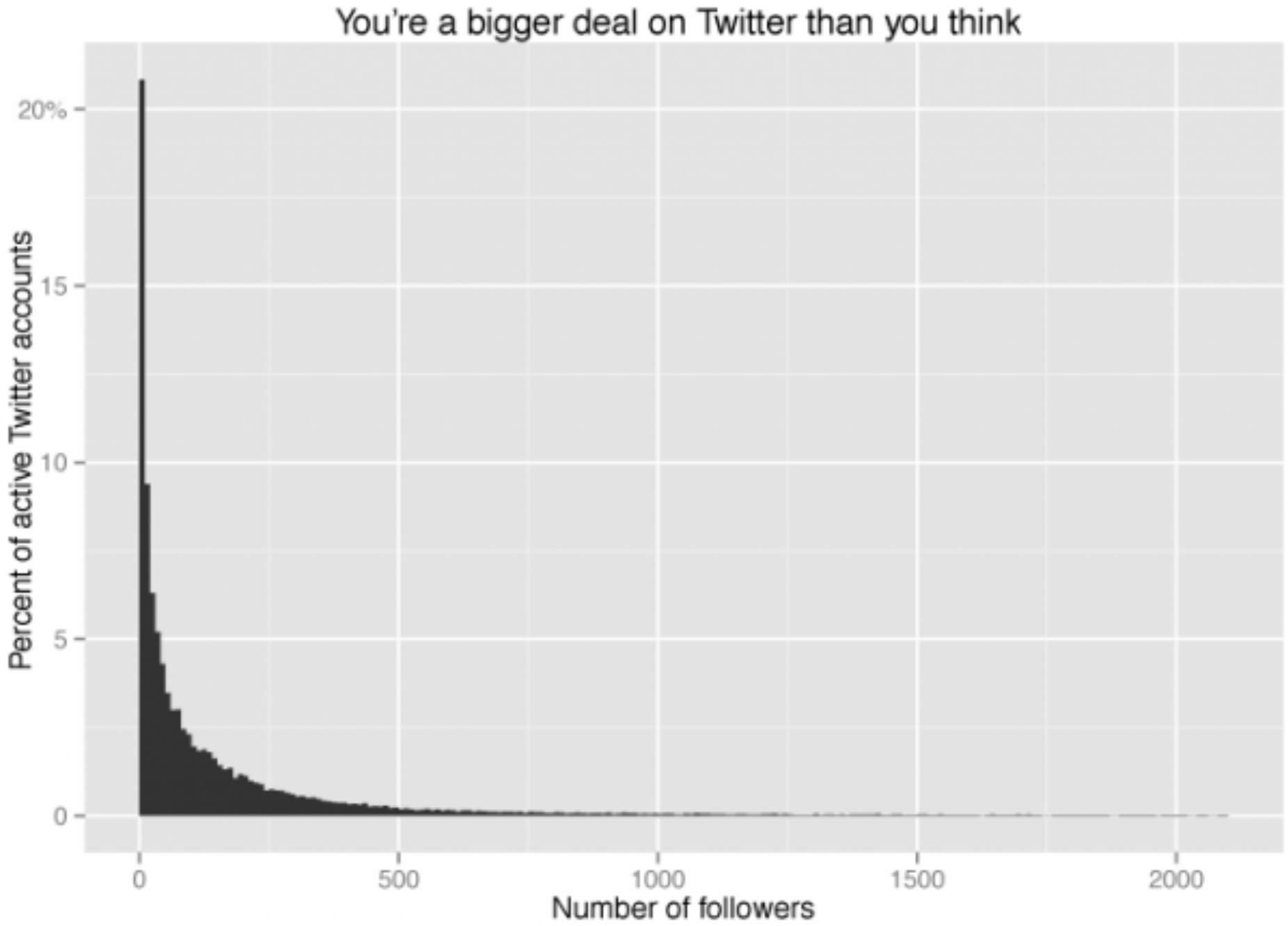


Figure 0 1 Population age structure diagrams for countries with rapid, slow, zero, and negative population growth rates. (Data from Population Reference Bureau)

Source: Thinglink

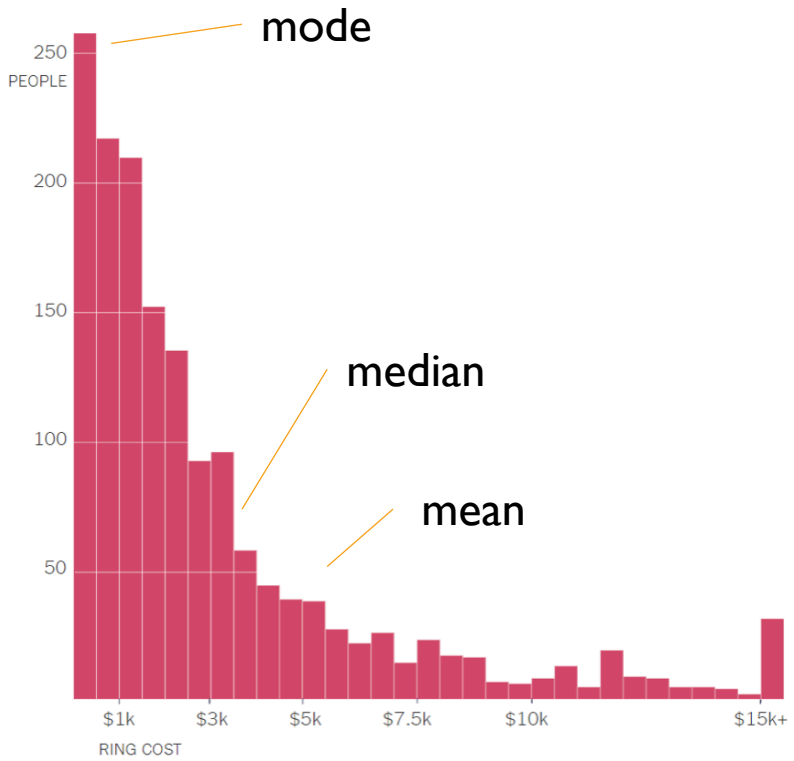


Source: BHIVE Social Media Labs

The New York Times

Engagement Ring Costs

How much do people spend on engagement rings?



Annual income	Share of income spent on an engagement ring
Less than $23,000	2.2 months (16.7%)
$23,000 to $40,000	2 weeks (4.0%)
$40,000 to $58,000	1.9 weeks (3.7%)
$58,000 to $90,000	1.6 weeks (3.1%)
More than $90,000	1.6 weeks (3.1%)

Source: National poll of 1,640 adults between Jan. 24 – 30, 2019 by Morning Consult

Learn about the article & graph here!

Interactives

- Population mountains
- Human terrain
- Summers are getting hotter

Outline

1. What is a data visualization?
2. Types of visualizations
3. What makes a good visualization?

matplotlib.org

MATPLOTLIB FOR VISUALIZATIONS

This is a great resource for information on aesthetics as well as code for each element (e.g., ticks, titles, colors)

https://python-graph-gallery.com/matplotlib/

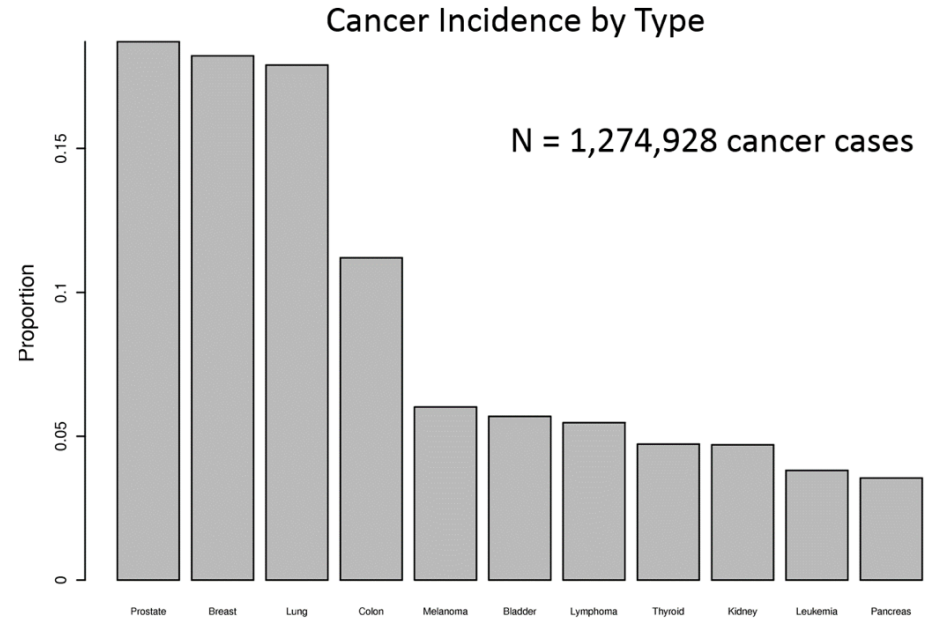
CHOICE OF TYPE OF VISUALIZATION IS IMPORTANT

Given your data type, what is the type of graph that is the best fit?

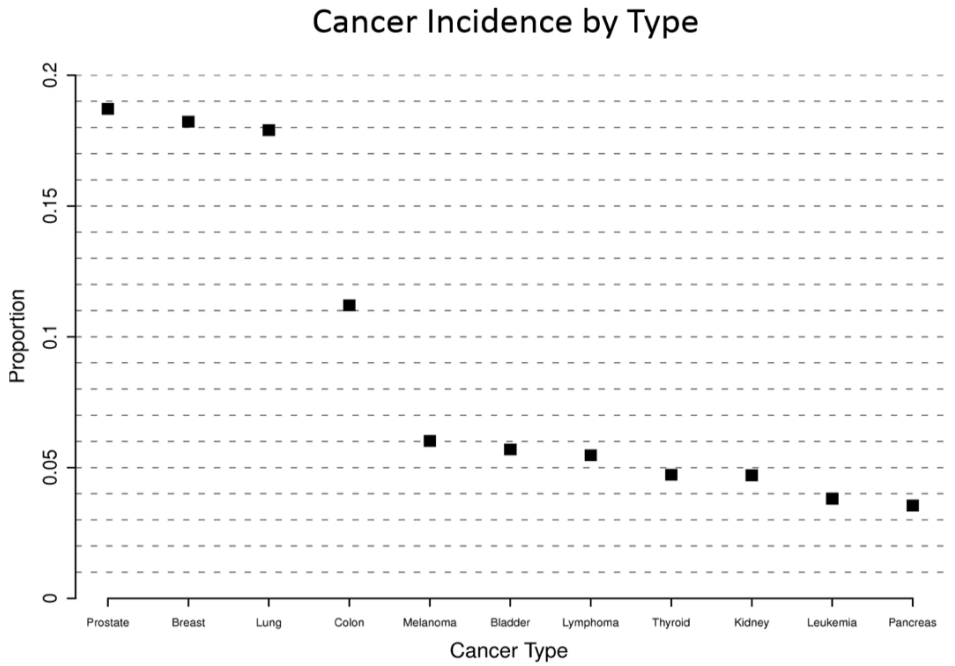
Sense-making

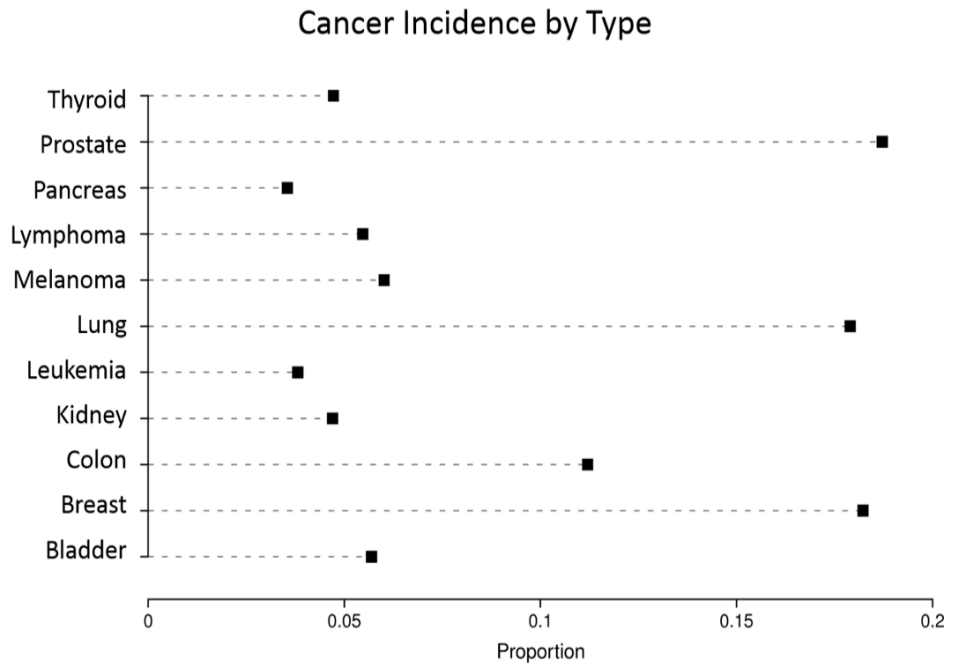
What story are you trying to tell? (What point are you trying to make?)

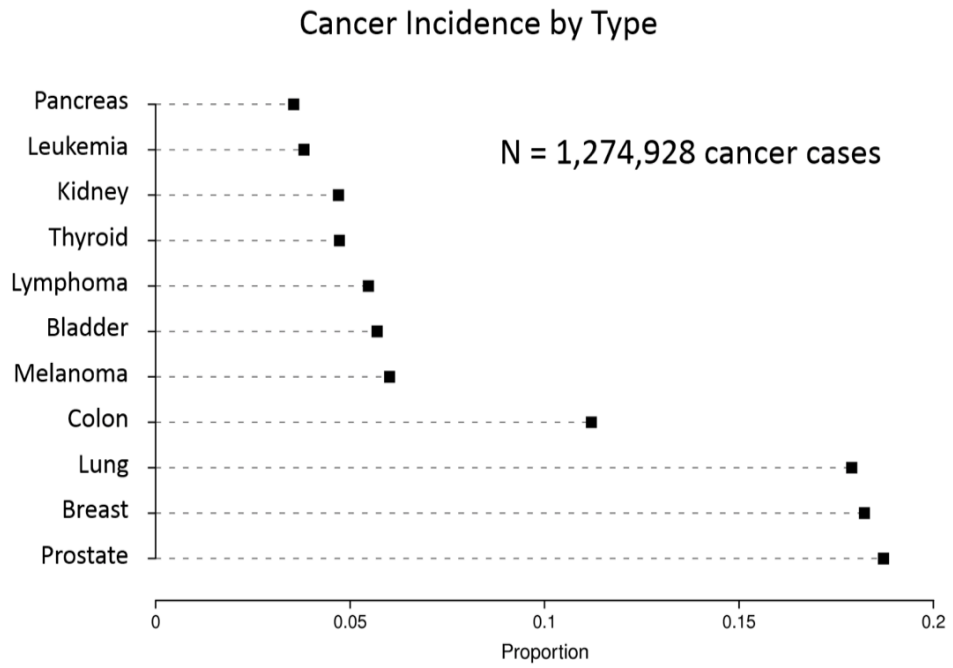
Communication

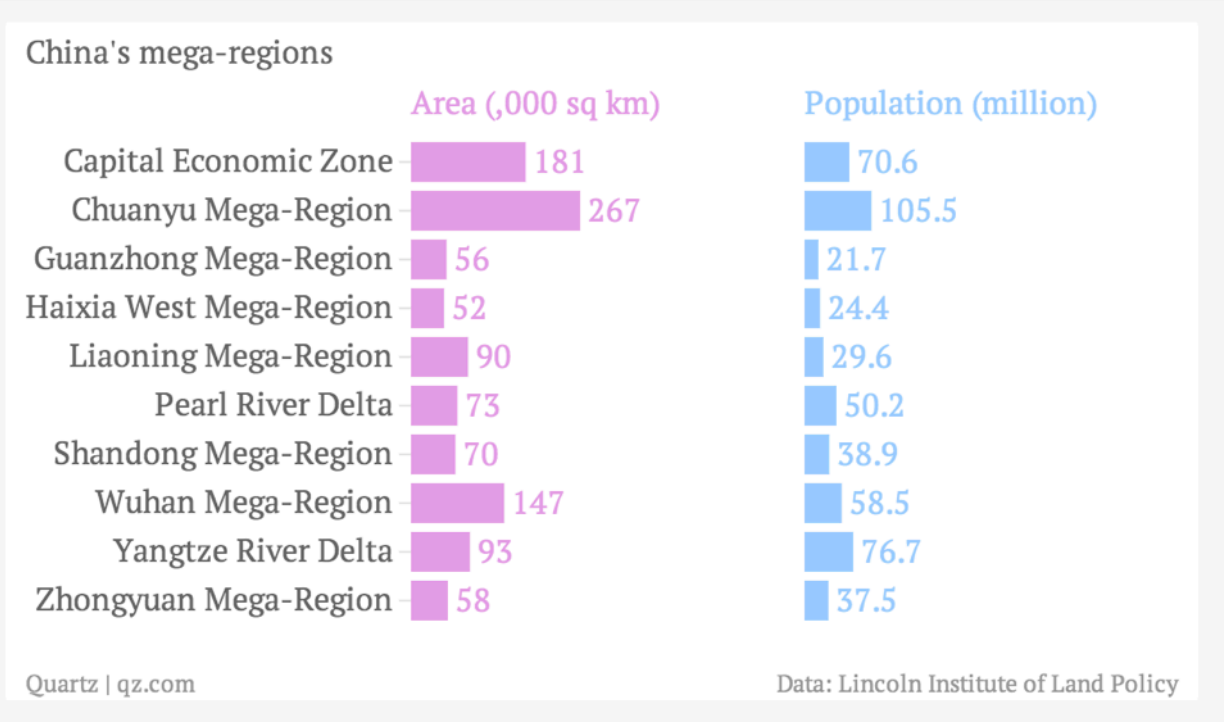

Cancer Incidence by Type
N = 1,274,928 cancer cases
Proportion
0
0.05
0.1
0.15
Prostate
Breast
Lung
Colon
Melanoma
Bladder
Lymphoma
Thyroid
Kidney
Leukemia
Pancreas

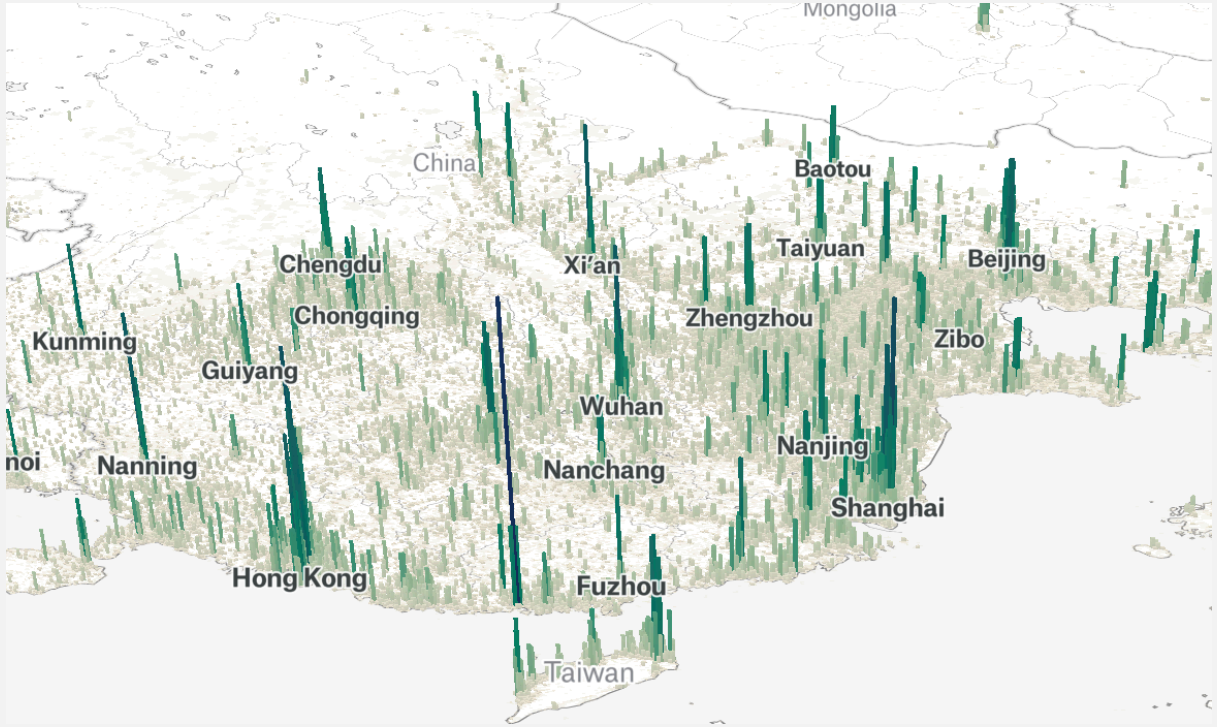
+


Cancer Incidence by Type
Proportion
0
0.05
0.1
0.15
0.2
Prostate
Breast
Lung
Colon
Melanoma
Bladder
Lymphoma
Thyroid
Kidney
Leukemia
Pancreas
Cancer Type


Cancer Incidence by Type
Thyroid
Prostate
Pancreas
Lymphoma
Melanoma
Lung
Leukemia
Kidney
Colon
Breast
Bladder
0
0.05
0.1
0.15
0.2
Proportion


Cancer Incidence by Type
N = 1,274,928 cancer cases
Pancreas
Leukemia
Kidney
Thyroid
Lymphoma
Bladder
Melanoma
Colon
Lung
Breast
Prostate
0
0.05
0.1
0.15
0.2
Proportion


China's mega-regions
Area (,000 sq km)
Population (million)
Capital Economic Zone
181
70.6
Chuanyu Mega-Region
267
105.5
Guanzhong Mega-Region
56
21.7
Haixia West Mega-Region
52
24.4
Liaoning Mega-Region
90
29.6
Pearl River Delta
73
50.2
Shandong Mega-Region
70
38.9
Wuhan Mega-Region
147
58.5
Yangtze River Delta
93
76.7
Zhongyuan Mega-Region
58
37.5
Quartz | qz.com
Data: Lincoln Institute of Land Policy


Mongolia
China
Baotou
Taiyuan
Beijing
Chengdu
Xi'an
Chongqing
Zhengzhou
Kunming
Zibo
Guiyang
Wuhan
Nanjing
Nanning
Nanchang
Shanghai
Hong Kong
Fuzhou
Taiwan

THE BASICS OF GOOD VISUALIZATIONS

1. Title that states the discovery and/or explains what's happening
   - Subtitle: If not in the title, exactly how to interpret what this is
2. Label everything, especially both axes, multiple variables, units, and sources
   - Ticks – Not too many, not too few!
3. All graphs should be able to stand on their own
   - I.e., they should make sense without surrounding text
4. Colors signify information.
   1. Should be used only to clarify or distinguish. Where possible use shapes and/or textures, too.
5. Consistency: Scale, color scheme, size
6. Simplicity!
   - Only include what's necessary. One major point per graphic. Be careful with multiple dimensions

Outline

1. What is a data visualization?
2. Types of visualizations
3. What makes a good visualization?