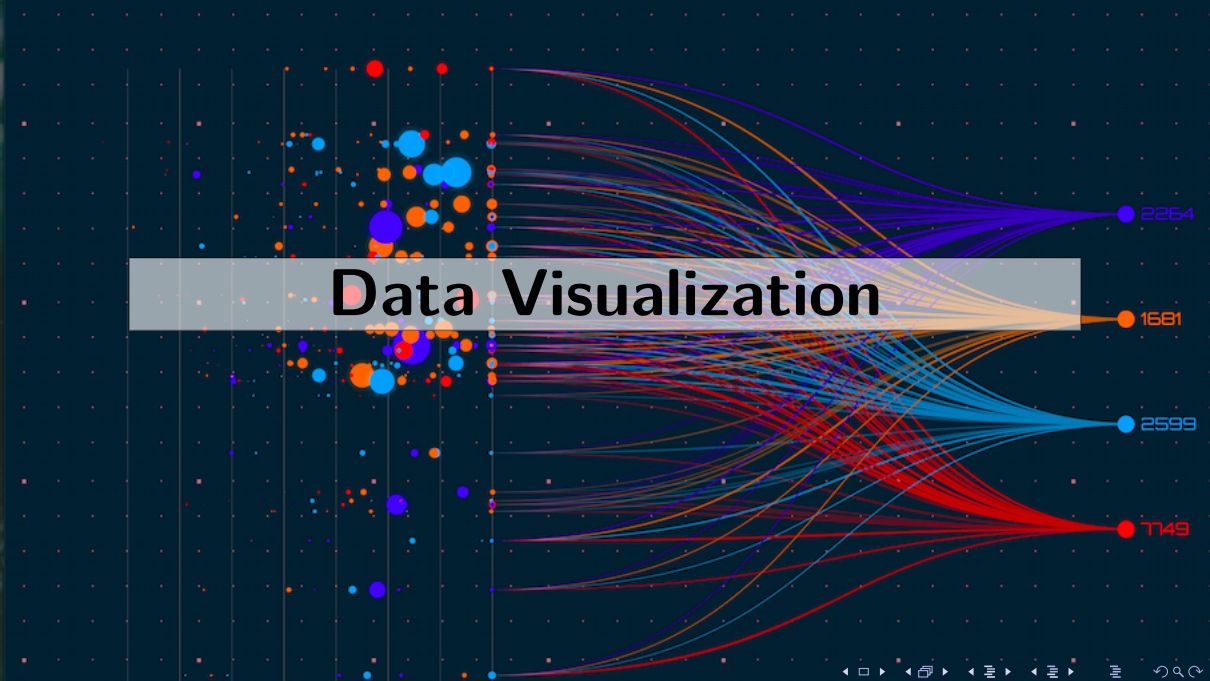

Data Visualization
2264
1681
2599
7749

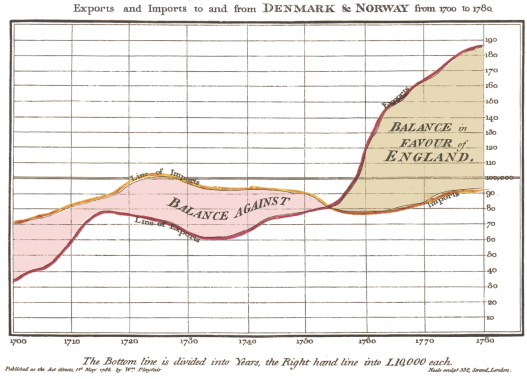
Today:

- Introduce some basic data visualizations;
- Introduce base R graphics API;
- Exemplars;
- Chart crimes.

Some history...

- Many ways to extract value from data:
  - descriptive statistics;
  - hypothesis testing;
  - simulation;
  - statistical modeling;
  - machine learning/AI;

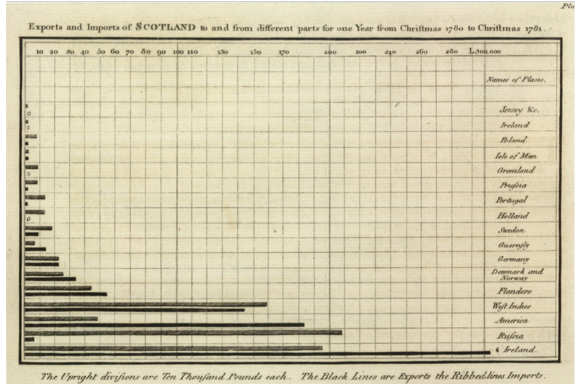
- One of the very oldest is **data visualization**:
  - William Playfair often cited as the inventor of line, bar, and pie charts, late 18th century.



Some history...

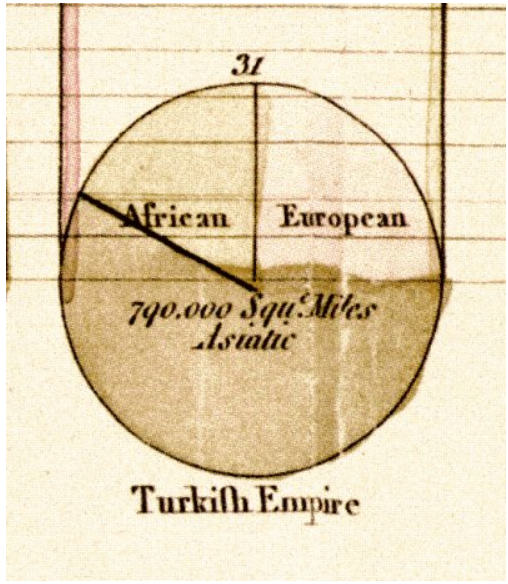
- Many ways to extract value from data:
  - descriptive statistics;
  - hypothesis testing;
  - simulation;
  - statistical modeling;
  - machine learning/AI;

- One of the very oldest is **data visualization**:
  - William Playfair often cited as the inventor of line, bar, and pie charts, late 18th century.



Some history...

- Many ways to extract value from data:
  - descriptive statistics;
  - hypothesis testing;
  - simulation;
  - statistical modeling;
  - machine learning/AI;
- One of the very oldest is **data visualization**:
  - William Playfair often cited as the inventor of line, bar, and pie charts, late 18th century.



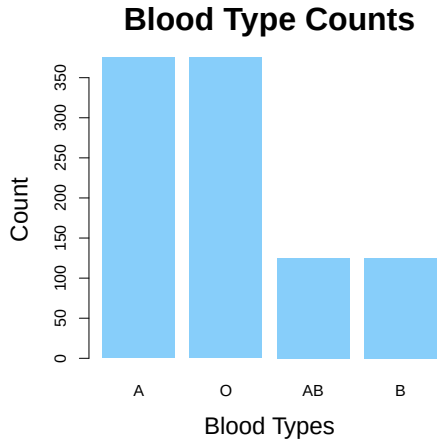
Goals:

- Characteristics of data visualization:
  - Data encoded as a visual object;
  - Goal of any data analysis is to help audience reason about the data – data viz is most reliant on human cognition for inference and meaning;
  - Form = substance;
  - Generally, necessarily low dimensional;

- Data visualization should (Tufte 1983):
  - Have message fit for audience purpose and communicate information;
  - Highlight the data;
  - Help the audience focus on the substance;
  - NOT distort the data or knowingly perpetuate a lie;
  - Put many numbers in a small, accessible space;
  - Enable the audience to grapple with large data sets coherently;
  - Encourage viewers to compare data effectively;
  - Allow investigation of both micro and macro level structure in the data;
  - integrate closely with additional statistical analysis.

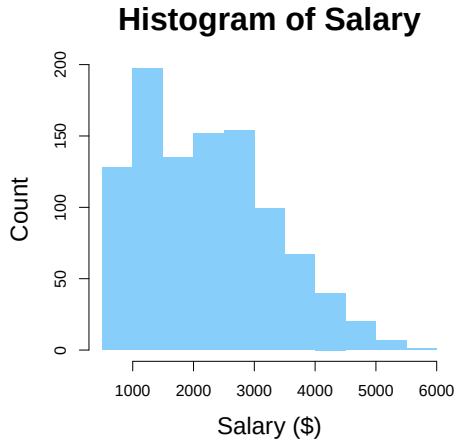
Some of my favorite types of graphics

- **Bar chart**;
  - Compare amount in fixed bins;
  - Used for **categorical** data where there is no order;
- **Histogram**;
  - Compare amount in constructed bins;
  - Used for **continuous** data;
- **Density plot**;
  - Smoothed comparison of amount in constructed bins;
  - Used for **continuous** data;
- **Scatterplot**;
  - Compare two variables;
  - Used for $(\mathbf{x}, \mathbf{y})$ data;



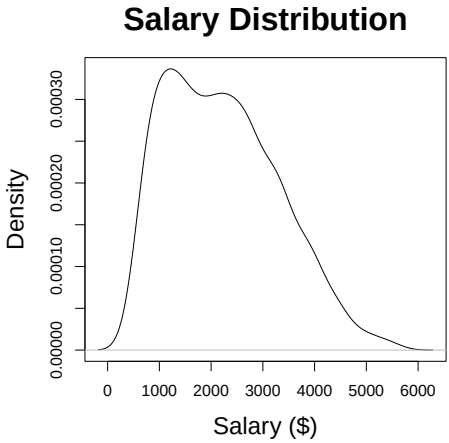
Some of my favorite types of graphics

- **Bar chart**;
  - Compare amount in fixed bins;
  - Used for **categorical** data where there is no order;
- **Histogram**;
  - Compare amount in constructed bins;
  - Used for **continuous** data;
- **Density plot**;
  - Smoothed comparison of amount in constructed bins;
  - Used for **continuous** data;
- **Scatterplot**;
  - Compare two variables;
  - Used for $(\mathbf{x}, \mathbf{y})$ data;



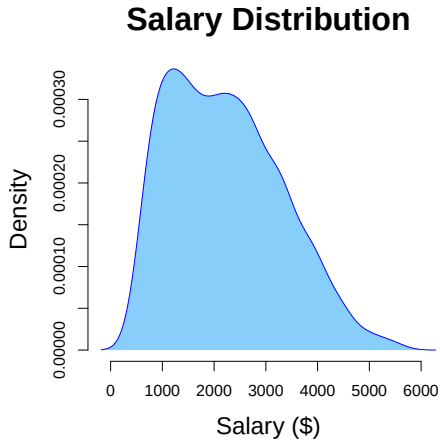
Some of my favorite types of graphics

- **Bar chart**;
  - Compare amount in fixed bins;
  - Used for **categorical** data where there is no order;
- **Histogram**;
  - Compare amount in constructed bins;
  - Used for **continuous** data;
- **Density plot**;
  - Smoothed comparison of amount in constructed bins;
  - Used for **continuous** data;
- **Scatterplot**;
  - Compare two variables;
  - Used for $(\mathbf{x}, \mathbf{y})$ data;



Some of my favorite types of graphics

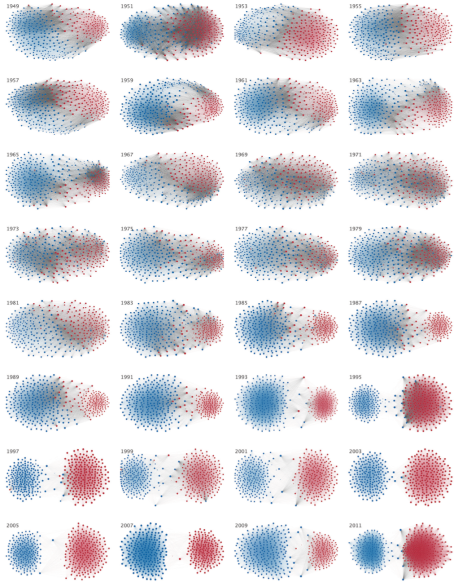
- **Bar chart**;
  - Compare amount in fixed bins;
  - Used for **categorical** data where there is no order;
- **Histogram**;
  - Compare amount in constructed bins;
  - Used for **continuous** data;
- **Density plot**;
  - Smoothed comparison of amount in constructed bins;
  - Used for **continuous** data;
- **Scatterplot**;
  - Compare two variables;
  - Used for $(\mathbf{x}, \mathbf{y})$ data;



Some of my favorite types of graphics

- **Bar chart**;
  - Compare amount in fixed bins;
  - Used for **categorical** data where there is no order;
- **Histogram**;
  - Compare amount in constructed bins;
  - Used for **continuous** data;
- **Density plot**;
  - Smoothed comparison of amount in constructed bins;
  - Used for **continuous** data;
- **Scatterplot**;
  - Compare two variables;
  - Used for $(\mathbf{x}, \mathbf{y})$ data;

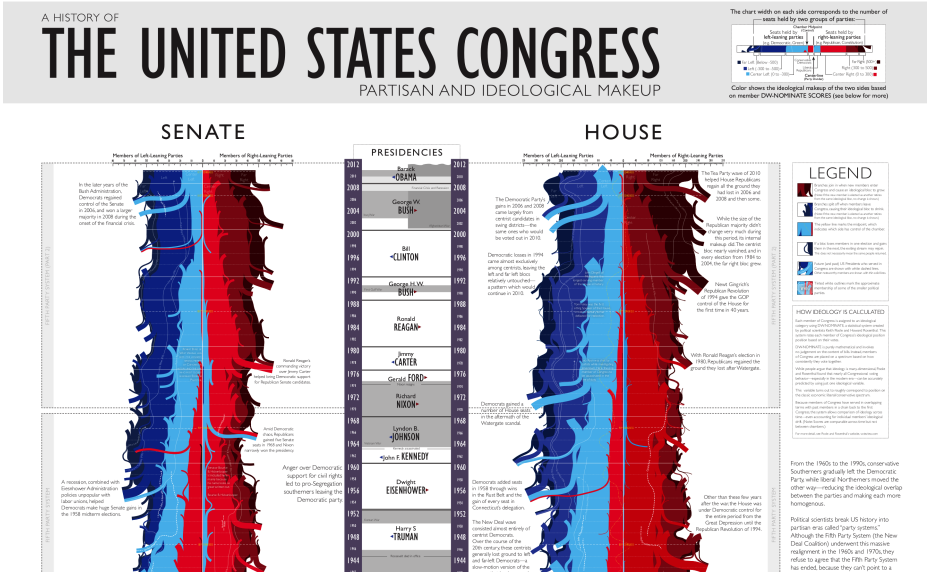
Congressional Partisanship



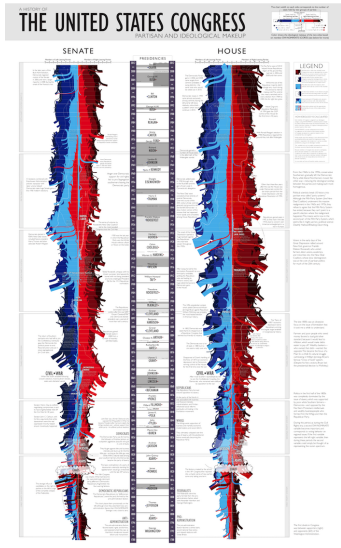
Congressional Partisanship: dimensions of information

- Number of legislators;
- Party identification;
- Legislator location in ‘ideological space’;
- Number of vote associations between legislators;
- Time in years.

Makeup of Congress



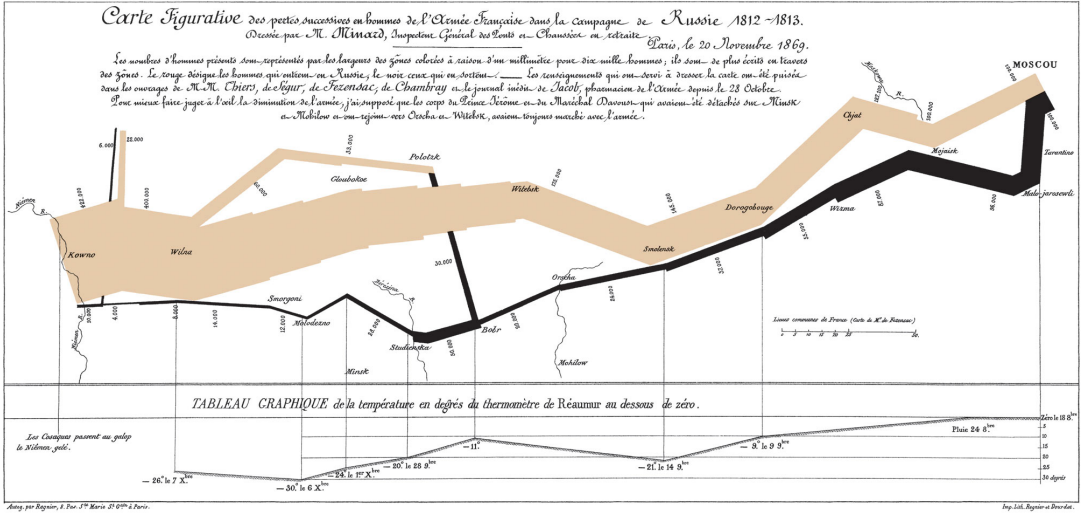
Makeup of Congress



Makeup of Congress: dimensions of information

- Number of legislators;
- Size of parties in congress;
- Size of congress itself;
- Size of intra-party ideological blocks;
- Block entry/exit;
- Time in years.

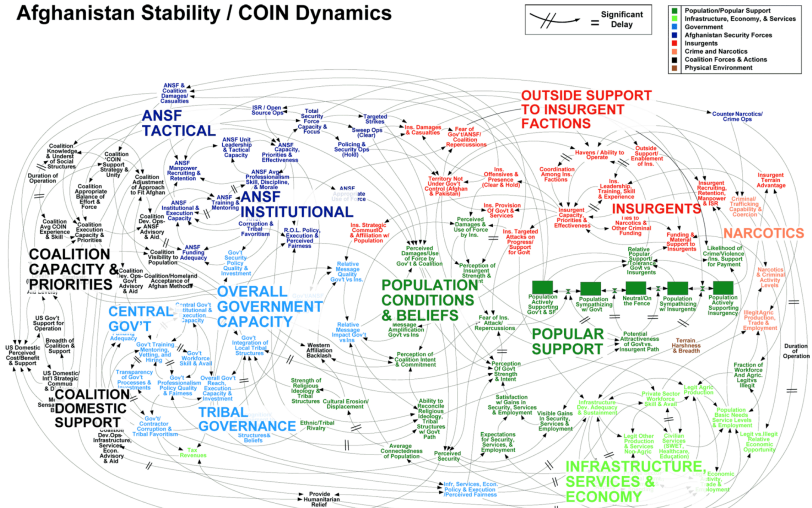
Napoleon's Progress



Napoleon's Progress: dimensions of information

- Distance traveled by French Army;
- Geographic location in terms of latitude and longitude;
- Number of soldiers capable of fighting in French Army – relative to beginning of campaign;
- French army advancing or retreating?
- Time of year;
- Temperature.

Dynamics in Afghanistan



Impossible to focus.

COVID Timeline

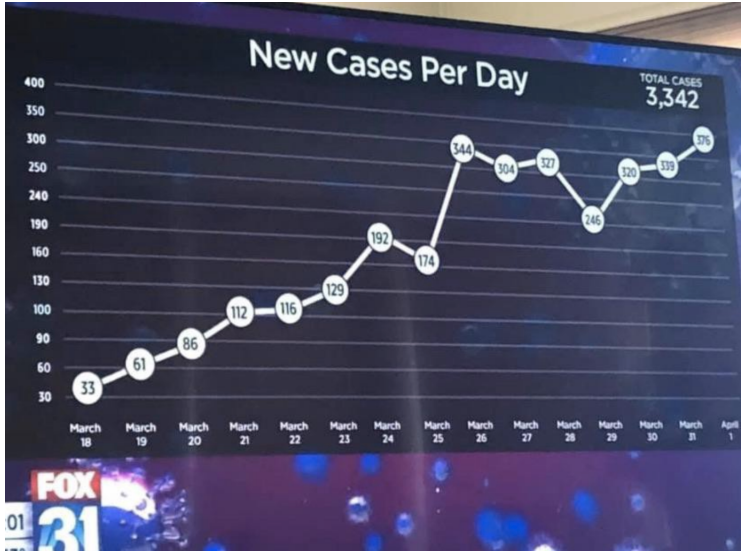


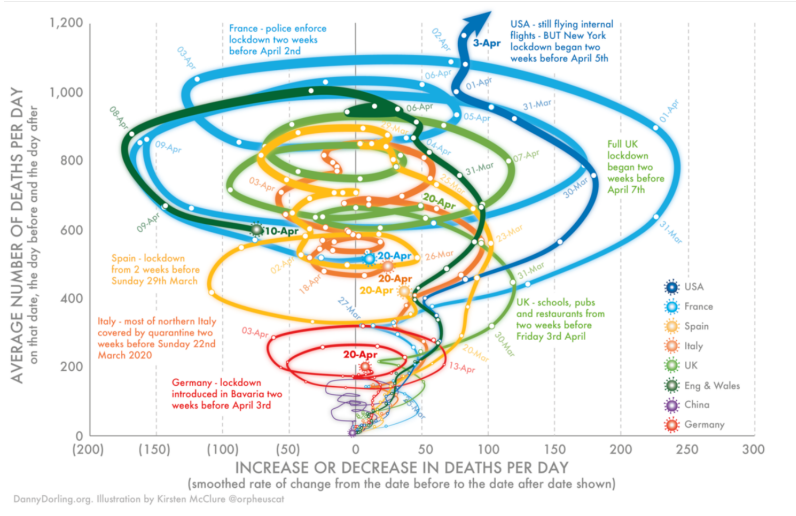
Chart Crime:

Non-linear y-axis.

COVID's Reach

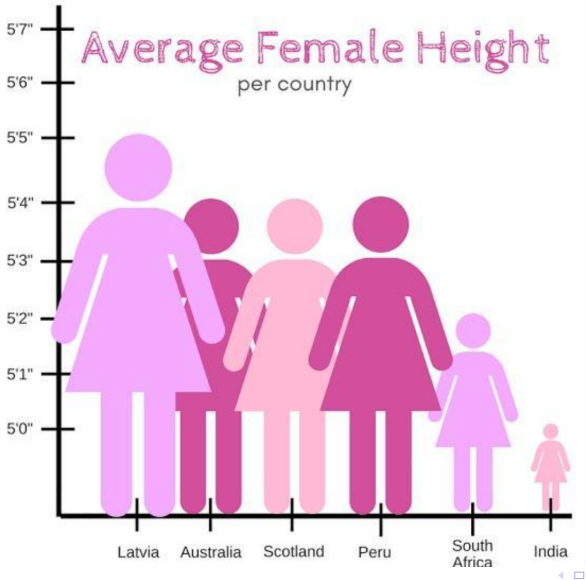
Underlying data is unrelated to graphic message.

WTF?!


France - police enforce lockdown two weeks before April 2nd
USA - still flying internal flights - BUT New York lockdown began two weeks before April 5th
Full UK lockdown began two weeks before April 7th
Spain - lockdown from 2 weeks before Sunday 29th March
Italy - most of northern Italy covered by quarantine two weeks before Sunday 22nd March 2020
Germany - lockdown introduced in Bavaria two weeks before April 3rd
UK - schools, pubs and restaurants from two weeks before Friday 3rd April
3-Apr
10-Apr
20-Apr
AVERAGE NUMBER OF DEATHS PER DAY
on that date, the day before and the day after
INCREASE OR DECREASE IN DEATHS PER DAY
(smoothed rate of change from the date before to the date after date shown)
USA
France
Spain
Italy
UK
Eng & Wales
China
Germany
DannyDorling.org, Illustration by Kirsten McClure @orpheuscat

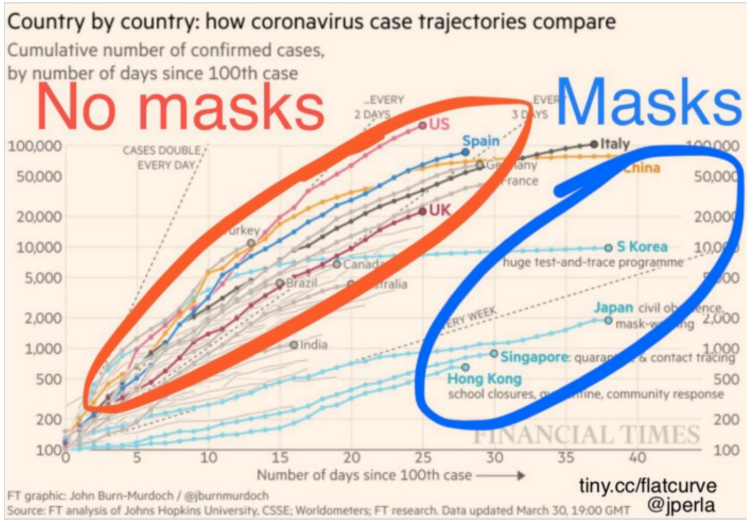
No clear message, extremely difficult to interpret.

Height Differences by Country

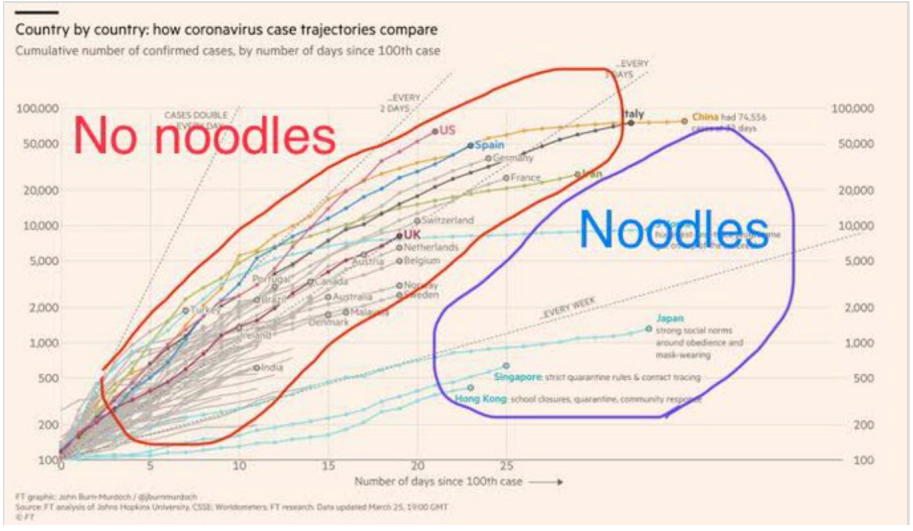


Truncated y-axis, misleading area comparison.

Causality?


Country by country: how coronavirus case trajectories compare
Cumulative number of confirmed cases,
by number of days since 100th case
No masks
Masks
...EVERY 2 DAYS
...EVERY 3 DAYS
CASES DOUBLE, EVERY DAY
US
Spain
Italy
France
China
UK
Turkey
Brazil
Canada
India
S Korea
huge test-and-trace programme
Japan civil obedience
mask-wearing
Singapore: quarantine & contact tracing
Hong Kong
school closures, quarantine, community response
100,000
50,000
20,000
10,000
5,000
2,000
1,000
500
200
100
0
5
10
15
20
25
30
35
40
Number of days since 100th case
FINANCIAL TIMES
tiny.cc/flatcurve
@jperla
FT graphic: John Burn-Murdoch / @jburnmurdoch
Source: FT analysis of Johns Hopkins University, CSSE; Worldometers; FT research. Data updated March 30, 19:00 GMT

Causality?


Country by country: how coronavirus case trajectories compare
Cumulative number of confirmed cases, by number of days since 100th case
No noodles
Noodles
CASES DOUBLE EVERY DAY
EVERY 2 DAYS
EVERY 3 DAYS
EVERY WEEK
US
Spain
Italy
China had 74,536 cases in 33 days
Germany
France
Iran
Switzerland
UK
Netherlands
Belgium
Austria
Portugal
Canada
Brazil
Australia
Norway
Sweden
Malaysia
Denmark
Ireland
Turkey
India
Japan
strong social norms around obedience and mask-wearing
Singapore strict quarantine rules & contract tracing
Hong Kong school closures, quarantine, community response
100,000
50,000
20,000
10,000
5,000
2,000
1,000
500
200
100
0
5
10
15
20
25
Number of days since 100th case
FT graphic: John Burn-Murdoch / @jburnmurdoch
Source: FT analysis of Johns Hopkins University, CSSE; Worldometers. FT research. Data updated March 25, 19:00 GMT
© FT

Asserting causality.

Sino-Indian Diplomacy

Misleading area comparison.

Why should we care?

A picture is worth a thousand table entries...