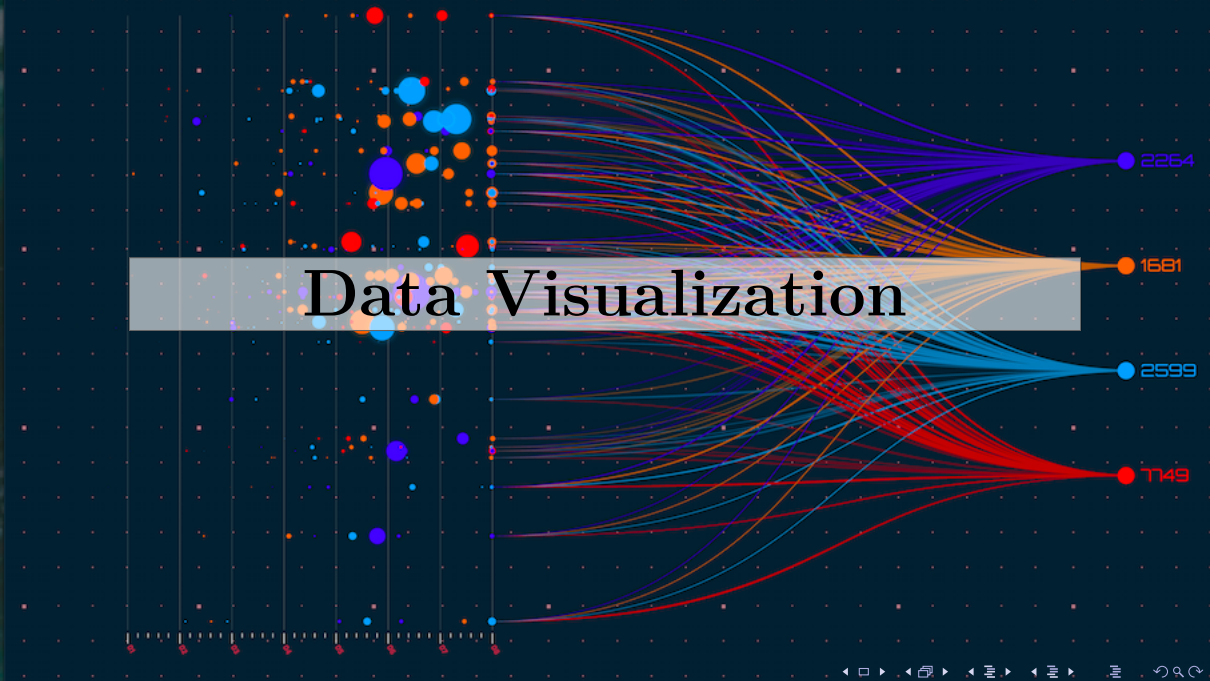
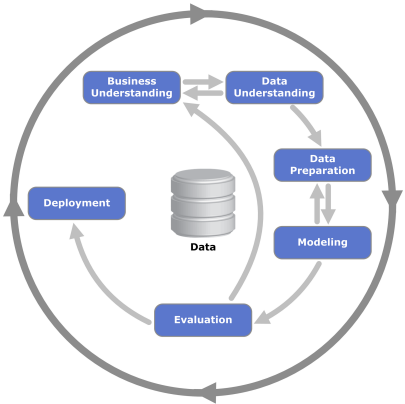


Data Visualization

Learning Outcomes



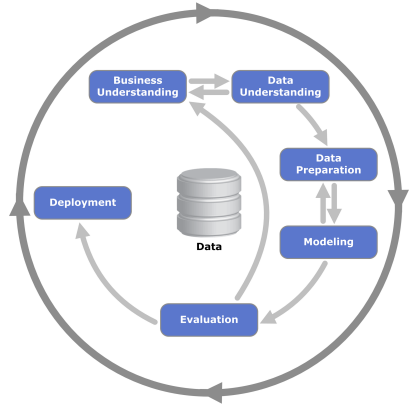
The CRISP-DM process.

- Define exploratory data analysis;
- List goals of visualizations;
- Interpret 10 commonly used visualizations;
- Offer critique of bad visualizations.

MSOE UNIVERSITY

Exploratory Data Analysis

- Exploratory data analysis (EDA) is the search for patterns and trends in a given data set;
- Goals:
  - Suggest interesting hypotheses;
  - Identify data processing mistakes/outliers;
  - Find violations of statistical assumptions;
  - Identify a potential set of features for ML.



EDA falls at the end of data prep and the beginning of modeling.

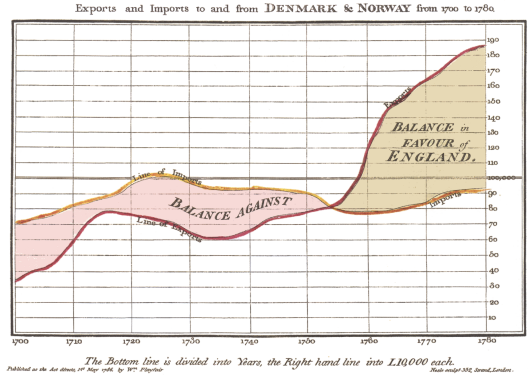
Confronting New Data

- Get a feel for the data by considering...
  - What is the data generating process? Who got the data, when, and why?
  - How big is it?
  - What do the fields mean?
  - Gut checks – look for familiar or interpretable records;
  - Summary statistics, e.g. Max, Min, Quartiles, Median, etc...
- ...and also by applying visualization!
  - A key part of EDA and beyond:
    - Get a feel for what your data really looks like. Identify trends over dataset that might not be obvious from summary statistics;
    - Detect errors, e.g. outliers, insufficient cleaning, coding mistakes, bad assumptions;
    - Communicate what you found – results only become actionable after they are shared.

Some history...

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

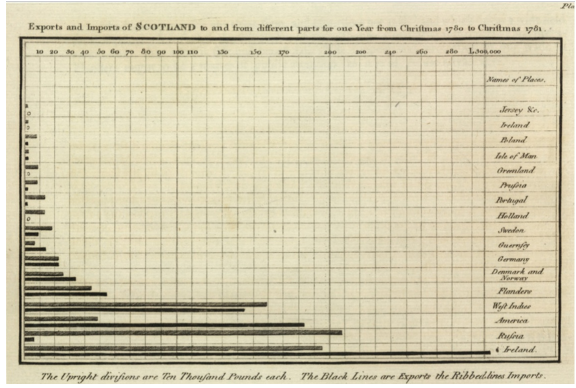
- One of the very oldest is **data visualization**:
  - William Playfair often cited as the inventor of line, bar, and pie charts, late 18th century.



Some history...

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

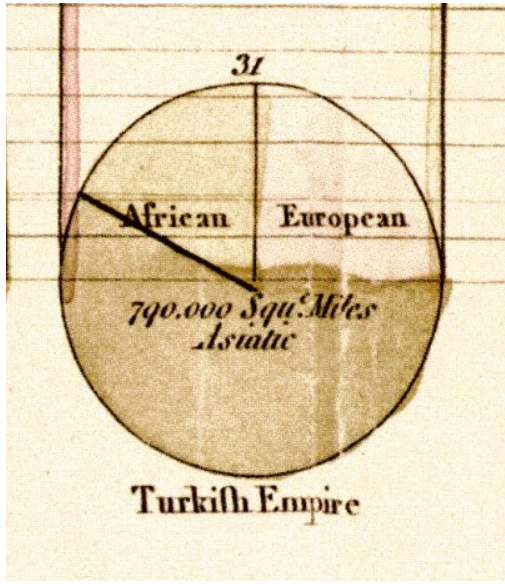
- One of the very oldest is **data visualization**:
  - William Playfair often cited as the inventor of line, bar, and pie charts, late 18th century.



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:

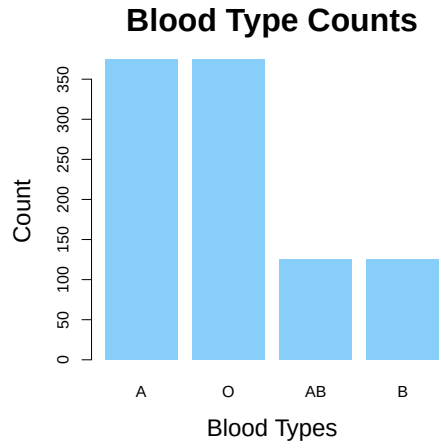
- Characteristics of data visualization:
  - Data encoded as a visual object;
  - Goal of any data analysis is to help audience reason about the data – data viz is most reliant on human cognition for inference and meaning;
  - Form = substance;
  - Generally, necessarily low dimensional;

- Data visualization should (Tufte 1983):
  - Have message fit for audience purpose and communicate information;
  - Highlight the data to 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;

- Libraries: Matplotlib, Seaborn, Pandas, Plotly.

Some of my favorite types of graphics – one variable

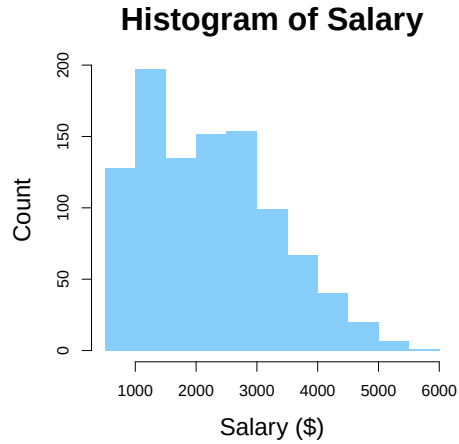
- **Bar chart**;
  - Compare amount in fixed bins (**categorical** data);
- **Histogram**;
  - Compare amount in constructed bins (**numerical** data);
- **Density plot**;
  - Smoothed amounts in constructed bins (**numerical** data);



A bar chart.

Some of my favorite types of graphics – one variable

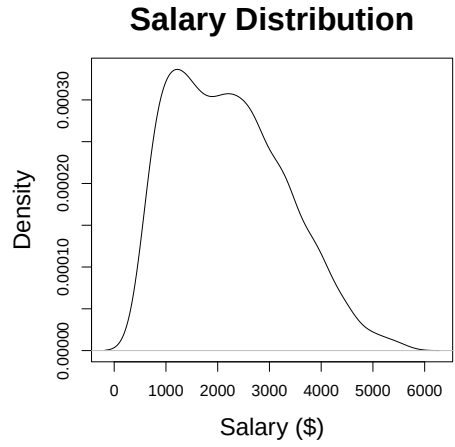
- **Bar chart**;
  - Compare amount in fixed bins (**categorical** data);
- **Histogram**;
  - Compare amount in constructed bins (**numerical** data);
- **Density plot**;
  - Smoothed amounts in constructed bins (**numerical** data);



A histogram.

Some of my favorite types of graphics – one variable

- **Bar chart**;
  - Compare amount in fixed bins (**categorical** data);
- **Histogram**;
  - Compare amount in constructed bins (**numerical** data);
- **Density plot**;
  - Smoothed amounts in constructed bins (**numerical** data);



A density.

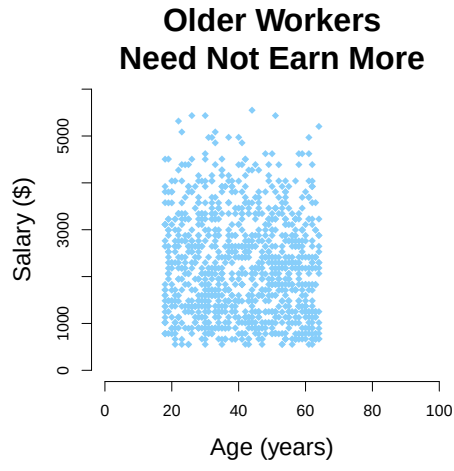
Some of my favorite types of graphics – one variable

- **Bar chart**;
  - Compare amount in fixed bins (**categorical** data);
- **Histogram**;
  - Compare amount in constructed bins (**numerical** data);
- **Density plot**;
  - Smoothed amounts in constructed bins (**numerical** data);

A nicer looking density.

Some of my favorite types of graphics – two variables

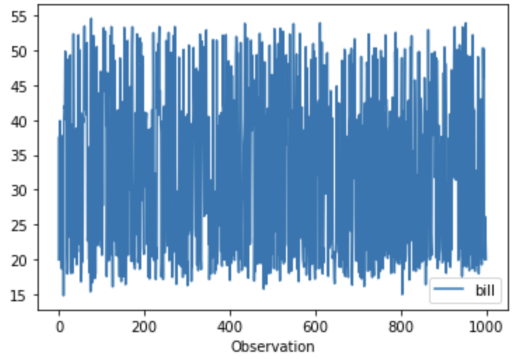
- **Scatterplot/line plot/scatterplot matrix**;
  - Compare two variables (**numerical/numerical** data);
- **Joint plot**;
  - Compare two variables (**numerical/numerical** data);
- **Bivariate Kernel Density Plot**;
  - Compare two variables (**numerical/numerical** data);
- **Boxplot/violin plot**;
  - Compare two variables (**categorical/numerical** data);
- **Heatmap**;
  - Compare two variables (**categorical/categorical** data);



A scatterplot.

Some of my favorite types of graphics – two variables

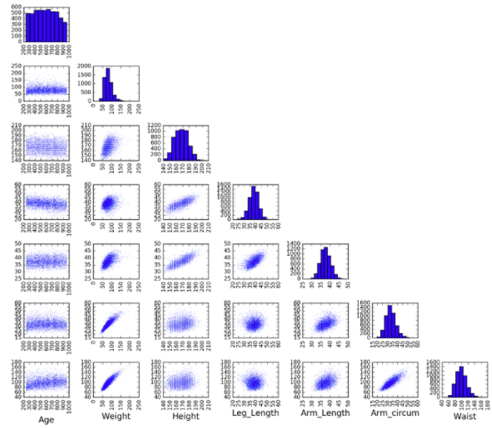
- **Scatterplot/line plot/scatterplot matrix**;
  - Compare two variables (**numerical/numerical** data);
- **Joint plot**;
  - Compare two variables (**numerical/numerical** data);
- **Bivariate Kernel Density Plot**;
  - Compare two variables (**numerical/numerical** data);
- **Boxplot/violin plot**;
  - Compare two variables (**categorical/numerical** data);
- **Heatmap**;
  - Compare two variables (**categorical/categorical** data);



A line plot.

Some of my favorite types of graphics – two variables

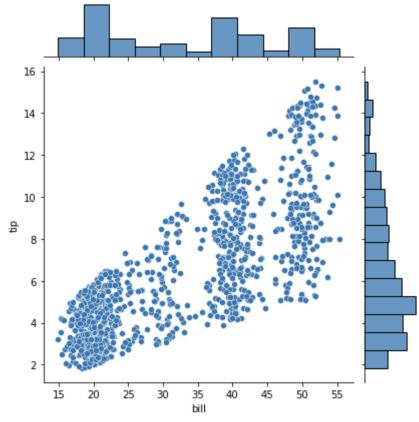
- **Scatterplot/line plot/scatterplot matrix**;
  - Compare two variables (**numerical/numerical** data);
- **Joint plot**;
  - Compare two variables (**numerical/numerical** data);
- **Bivariate Kernel Density Plot**;
  - Compare two variables (**numerical/numerical** data);
- **Boxplot/violin plot**;
  - Compare two variables (**categorical/numerical** data);
- **Heatmap**;
  - Compare two variables (**categorical/categorical** data);



A scatterplot matrix.

Some of my favorite types of graphics – two variables

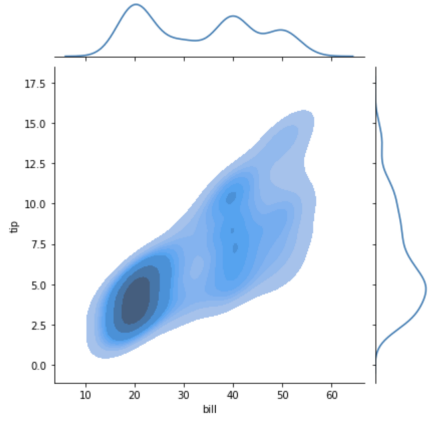
- **Scatterplot/line plot/scatterplot matrix**;
  - Compare two variables (**numerical/numerical** data);
- **Joint plot**;
  - Compare two variables (**numerical/numerical** data);
- **Bivariate Kernel Density Plot**;
  - Compare two variables (**numerical/numerical** data);
- **Boxplot/violin plot**;
  - Compare two variables (**categorical/numerical** data);
- **Heatmap**;
  - Compare two variables (**categorical/categorical** data);



A joint plot.

Some of my favorite types of graphics – two variables

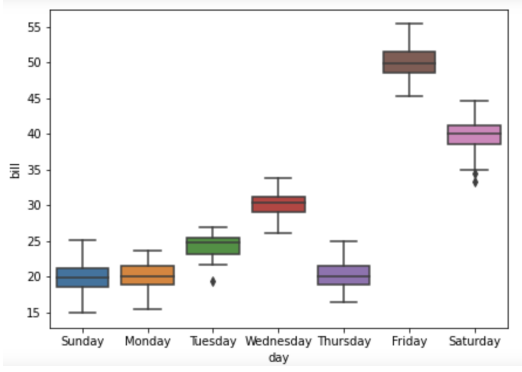
- **Scatterplot/line plot/scatterplot matrix**;
  - Compare two variables (**numerical/numerical** data);
- **Joint plot**;
  - Compare two variables (**numerical/numerical** data);
- **Bivariate Kernel Density Plot**;
  - Compare two variables (**numerical/numerical** data);
- **Boxplot/violin plot**;
  - Compare two variables (**categorical/numerical** data);
- **Heatmap**;
  - Compare two variables (**categorical/categorical** data);



A bivariate kernel density plot.

Some of my favorite types of graphics – two variables

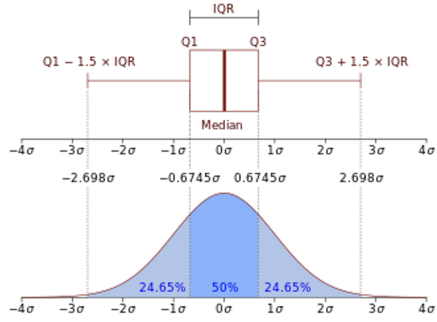
- **Scatterplot/line plot/scatterplot matrix**;
  - Compare two variables (**numerical/numerical** data);
- **Joint plot**;
  - Compare two variables (**numerical/numerical** data);
- **Bivariate Kernel Density Plot**;
  - Compare two variables (**numerical/numerical** data);
- **Boxplot/violin plot**;
  - Compare two variables (**categorical/numerical** data);
- **Heatmap**;
  - Compare two variables (**categorical/categorical** data);



A box plot.

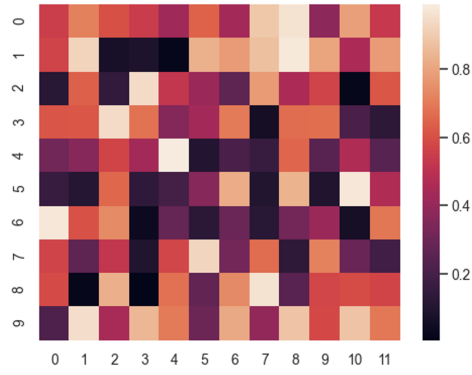
Some of my favorite types of graphics – two variables

- **Scatterplot/line plot/scatterplot matrix**;
  - Compare two variables (**numerical/numerical** data);
- **Joint plot**;
  - Compare two variables (**numerical/numerical** data);
- **Bivariate Kernel Density Plot**;
  - Compare two variables (**numerical/numerical** data);
- **Boxplot/violin plot**;
  - Compare two variables (**categorical/numerical** data);
- **Heatmap**;
  - Compare two variables (**categorical/categorical** data);



Some of my favorite types of graphics – two variables

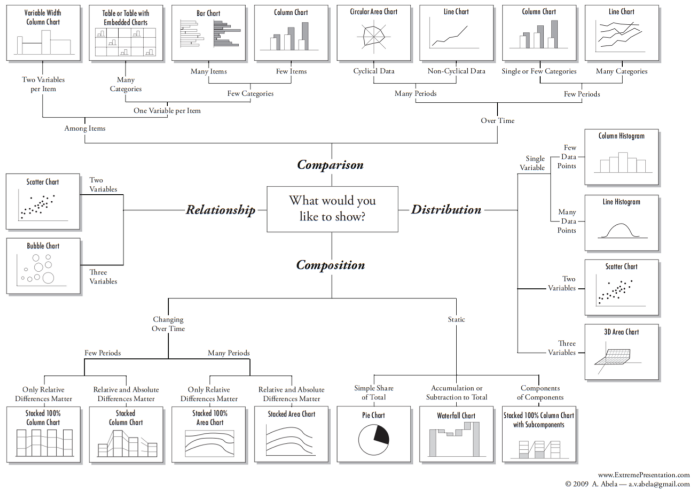
- **Scatterplot/line plot/scatterplot matrix**;
  - Compare two variables (**numerical/numerical** data);
- **Joint plot**;
  - Compare two variables (**numerical/numerical** data);
- **Bivariate Kernel Density Plot**;
  - Compare two variables (**numerical/numerical** data);
- **Boxplot/violin plot**;
  - Compare two variables (**categorical/numerical** data);
- **Heatmap**;
  - Compare two variables (**categorical/categorical** data);



And finally a heatmap.

MSOE UNIVERSITY

What kinds of charts are there?

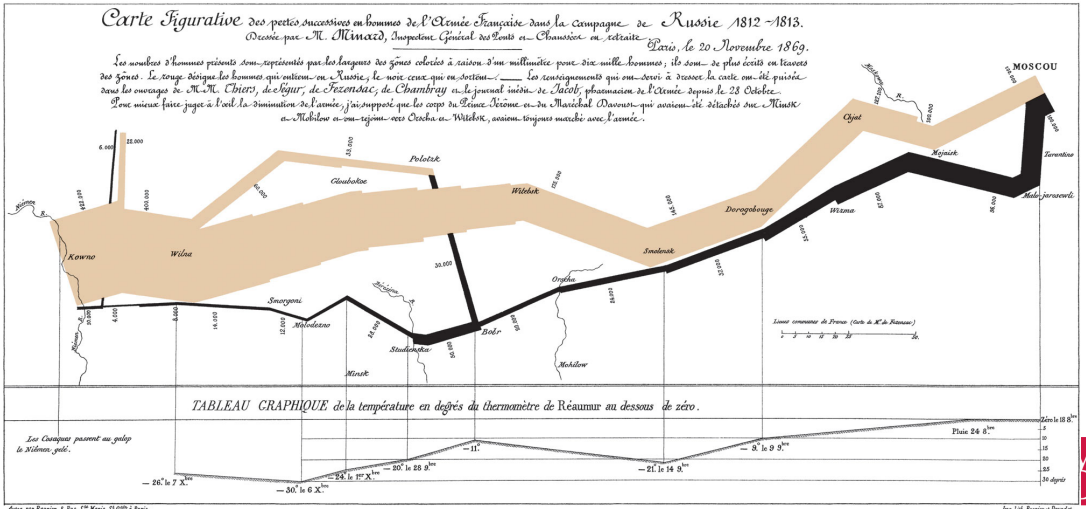


Python methods for all these...

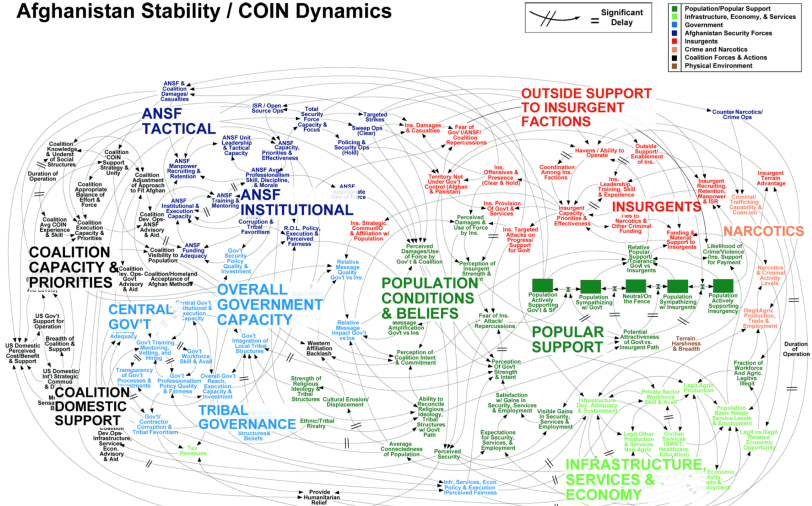
- Bar Charts:
  - Seaborn: .countplot();
  - Pandas: df.value_counts().plot_bar();
- Histograms:
  - Matplotlob: .hist();
  - Seaborn: .distplot();
  - Pandas: df.plot.hist();
- Densities:
  - Seaborn: .kdeplot();
  - Pandas: df.plot.kde();
- Scatterplots:
  - Matplotlib: .scatter();
  - Seaborn: .scatterplot(), .pairplot();
  - Pandas: df.plot.scatter();
- Line plot:
  - Matplotlib: .plot();
  - Seaborn: .lineplot();
  - Pandas: df.plot.line();
- Joint plot:
  - Seaborn: .jointplot();
- Bivariate Kernel Density plot:
  - Seaborn: .jointplot();
- Box plot:
  - Seaborn: .boxplot();
  - Pandas: df.boxplot();
- Violin plot:
  - Seaborn: .violinplot();
- Heatmap:
  - Seaborn: .heatmap();

Some examples...

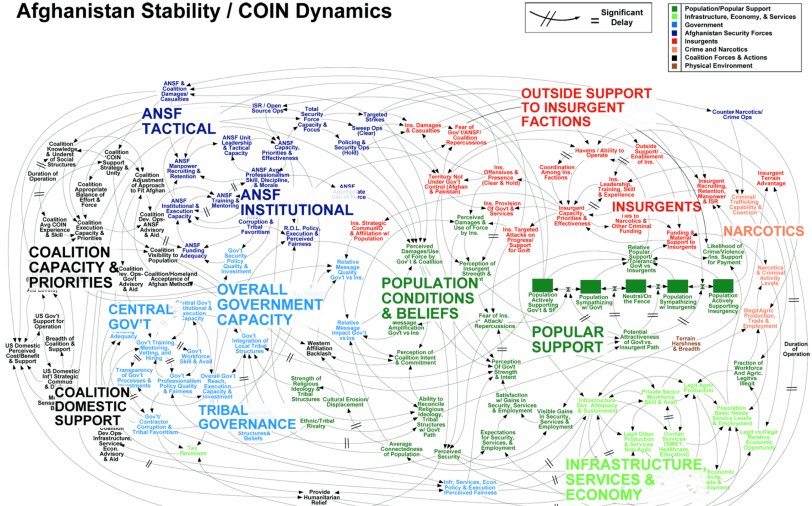
Napoleon's Progress – an exemplar?



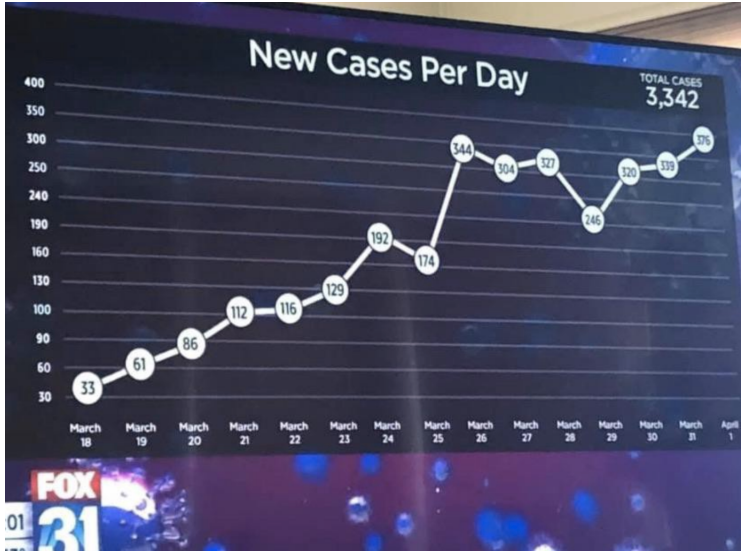
Dynamics in Afghanistan



Dynamics in Afghanistan – Impossible to focus.



COVID Timeline



COVID Timeline – Non-linear y-axis.

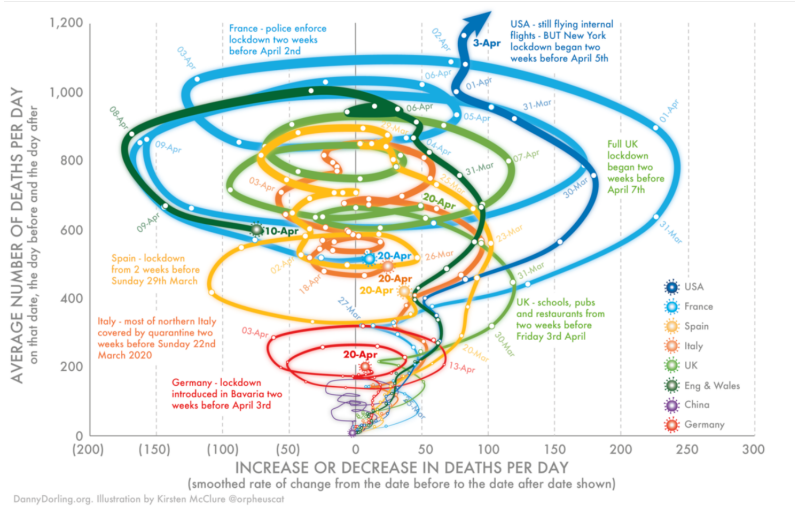
COVID's Reach

MSOE UNIVERSITY

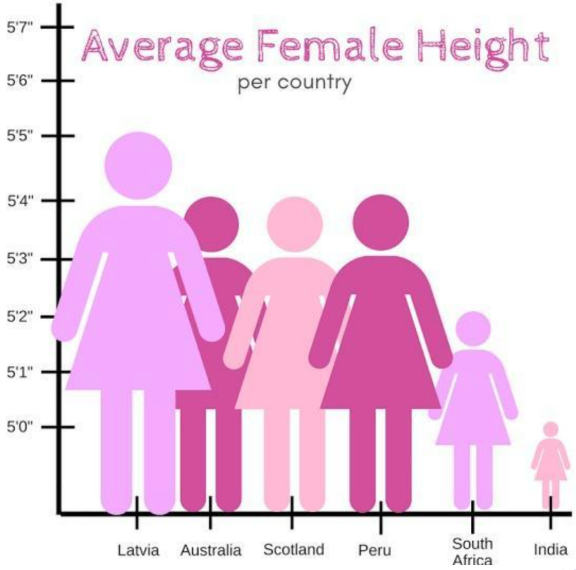
COVID's Reach – Underlying data is unrelated to graphic message.

MSOE UNIVERSITY

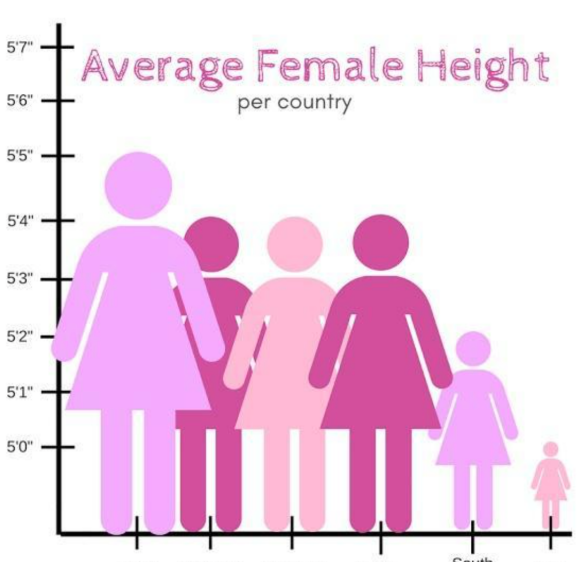
WTF?!


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)
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
USA
France
Spain
Italy
UK
Eng & Wales
China
Germany
DannyDorling.org, Illustration by Kirsten McClure @orpheuscat

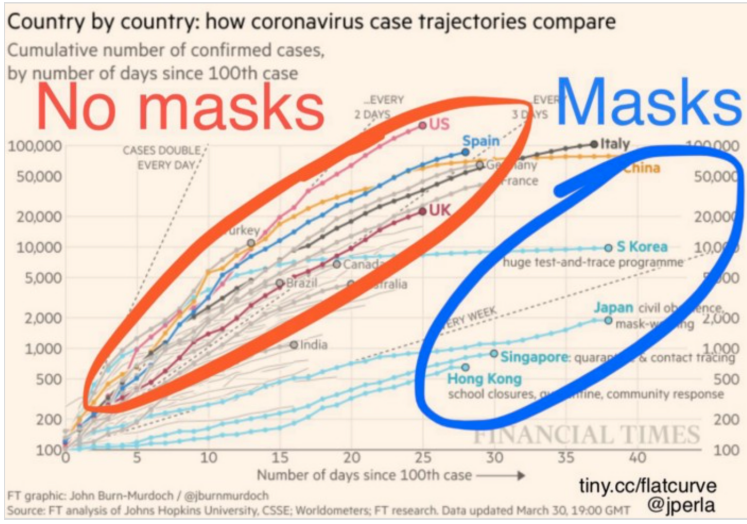
Height Differences by Country



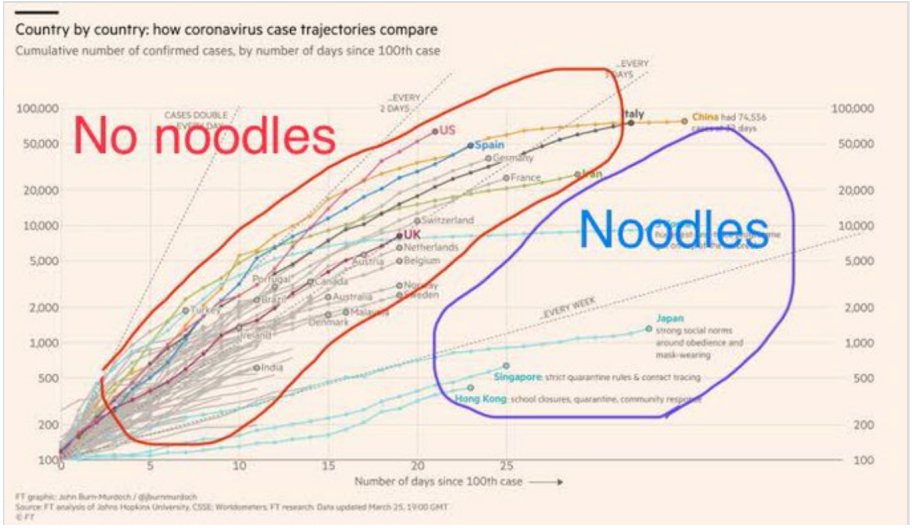
Height Differences by Country – Truncated y-axis, misleading area comparison.



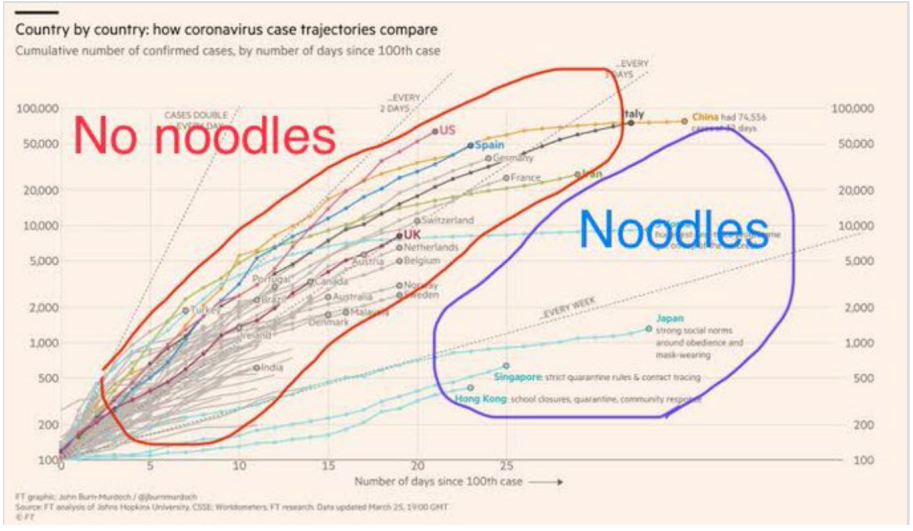
Causality?


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
China
France
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 53 days
Germany
France
Iran
Switzerland
UK
Netherlands
Belgium
Portugal
Canada
Austria
Brazil
Australia
Norway
Sweden
Turkey
Malaysia
Denmark
Ireland
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

Causality – Falsely asserting causality.



Sino-Indian Diplomacy

Sino-Indian Diplomacy – Misleading area comparison.