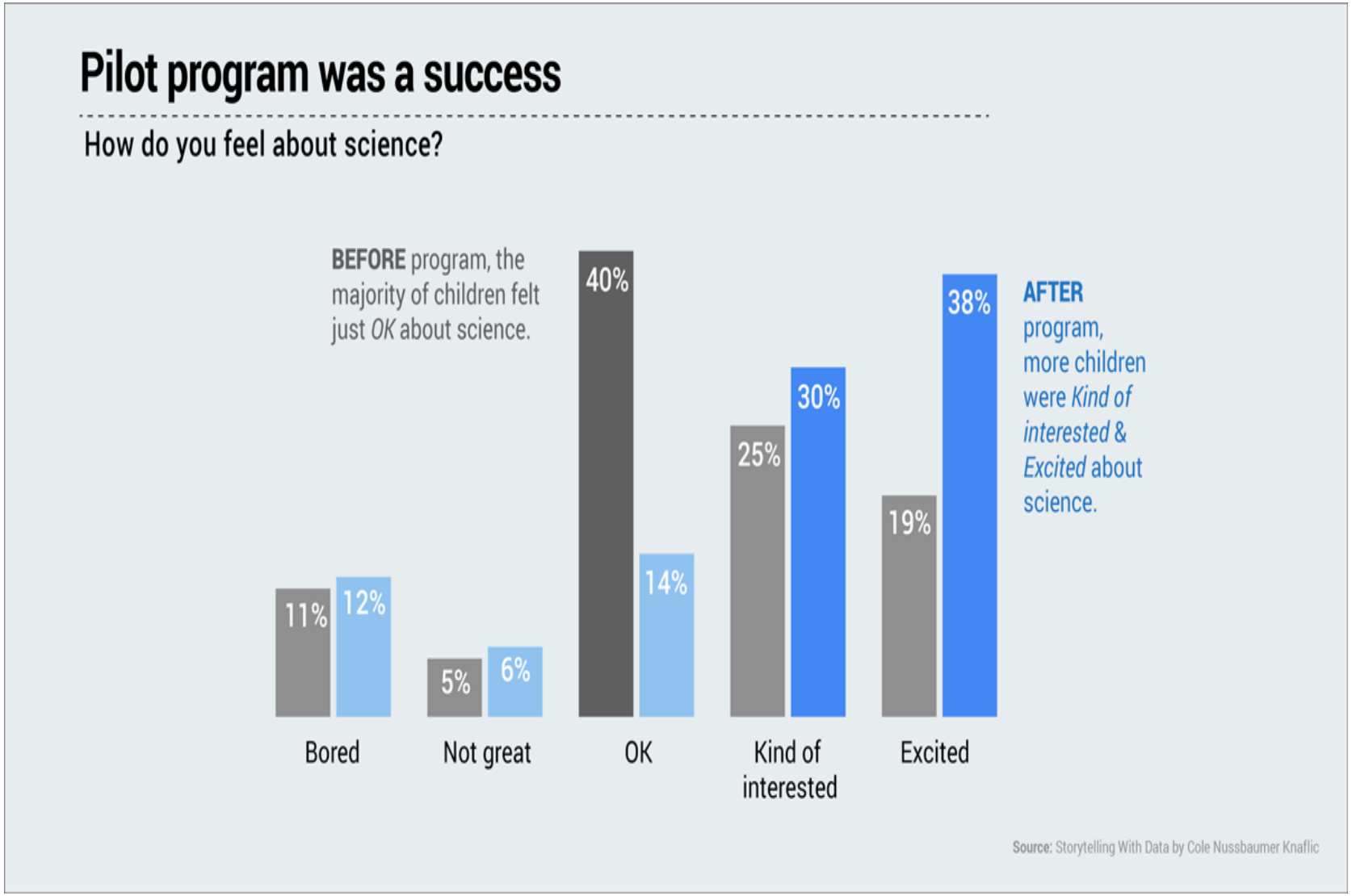
Data Visualisation
Lecture 5 – Visualising Relationships

Dr. Cathy Ennis

Learning Outcomes Week 5

- Design effective Visualisations based on principles from perceptual psychology, cognitive science, graphic design and visual art
- Create and deploy successful data visualisations using leading software tools
- Demonstrate an understanding how visualisation is used in date journalism to communicate complex ideas and stories
- Demonstrate understanding how visualisation is used in story telling

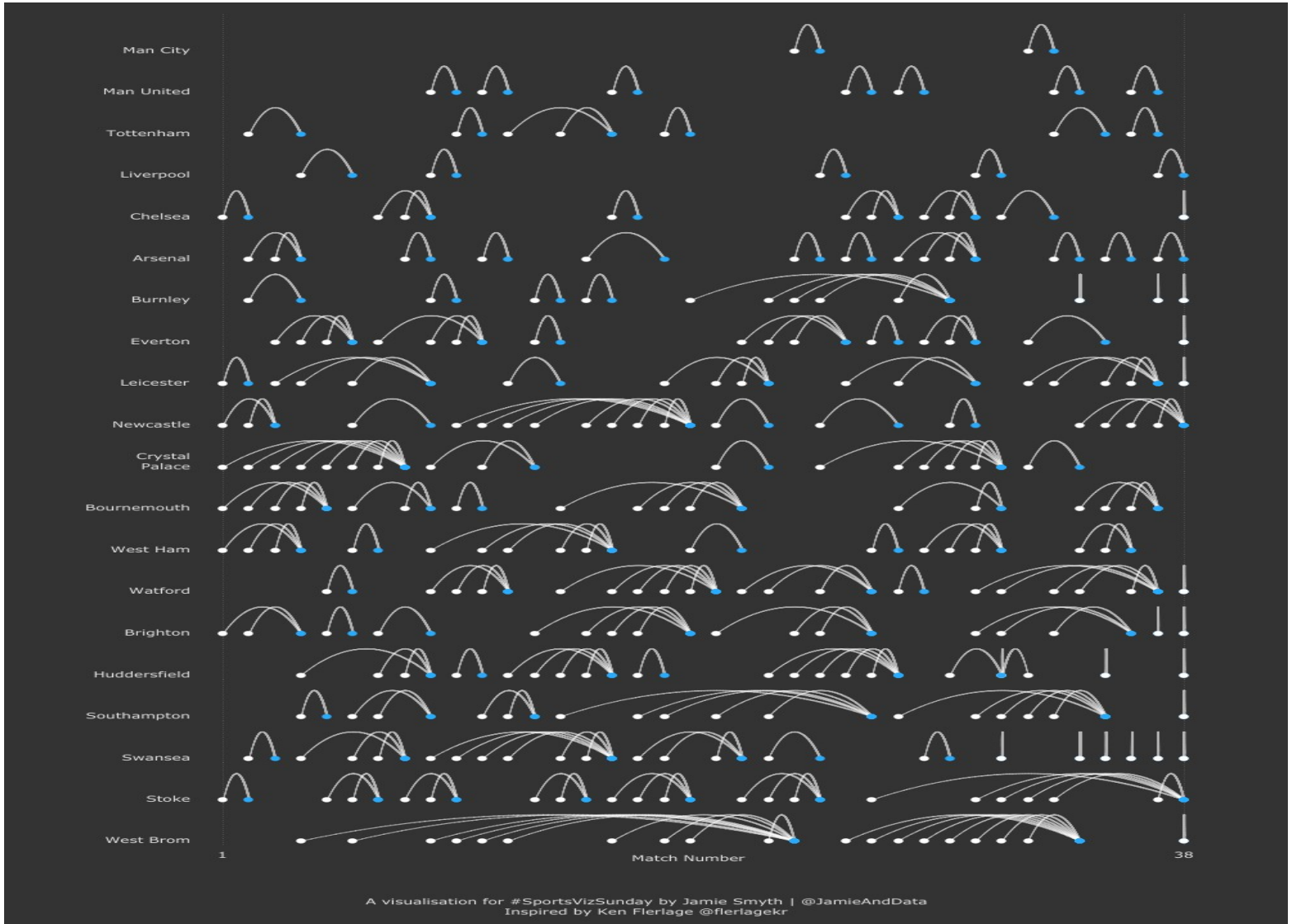
Visualisation of the Week



Visualisation Discussion Of The Week

https://public.tableau.com/en-us/s/gallery/english-premier-league-bouncebackability?gallery=votd

Visualisation Discussion Of The Week



(Un)Visualisation of the Week

Overview

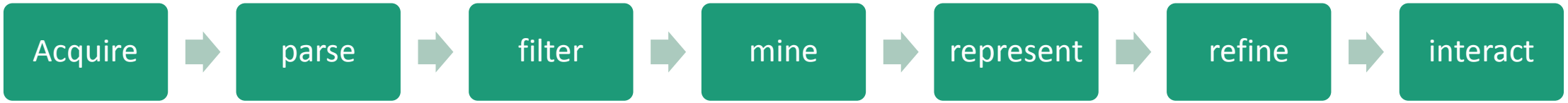
- Visualisation workflow

- Relationships
  - What relationships to look for: **Correlation**
  - Comparing variables: **Scatter plots, Line + Column**
  - Exploring more variables: **Bubble plots, 3D Scatter plots**
  - Exploring even more variables: **Scatter plot matrix, XY Heat Maps, Parallel Coordinates**

VISUALISATION WORKFLOW

Visualisation

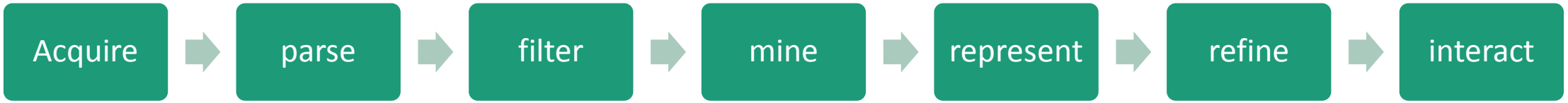
- Ben Fry's visualisation process - Designer



"Visualizing Data: Exploring and Explaining Data with the Processing Environment", B. Fry, O'Reilly Media, 2007

Visualisation

- Ben Fry's visualisation process - Designer



- Visualisation process – Viewer

"Visualizing Data: Exploring and Explaining Data with the Processing Environment", B. Fry, O'Reilly Media, 2007

Visualisation Workflow

Stage 1	Stage 2	Stage 3	Stage 4
• Formulating your Brief	• Working with data	• Establishing your **editorial thinking**	• Developing your design solution

Data Visualisation: A Handbook for Data Driven Design. Andy Kirk

Visualisation Workflow - Editorial Thinking

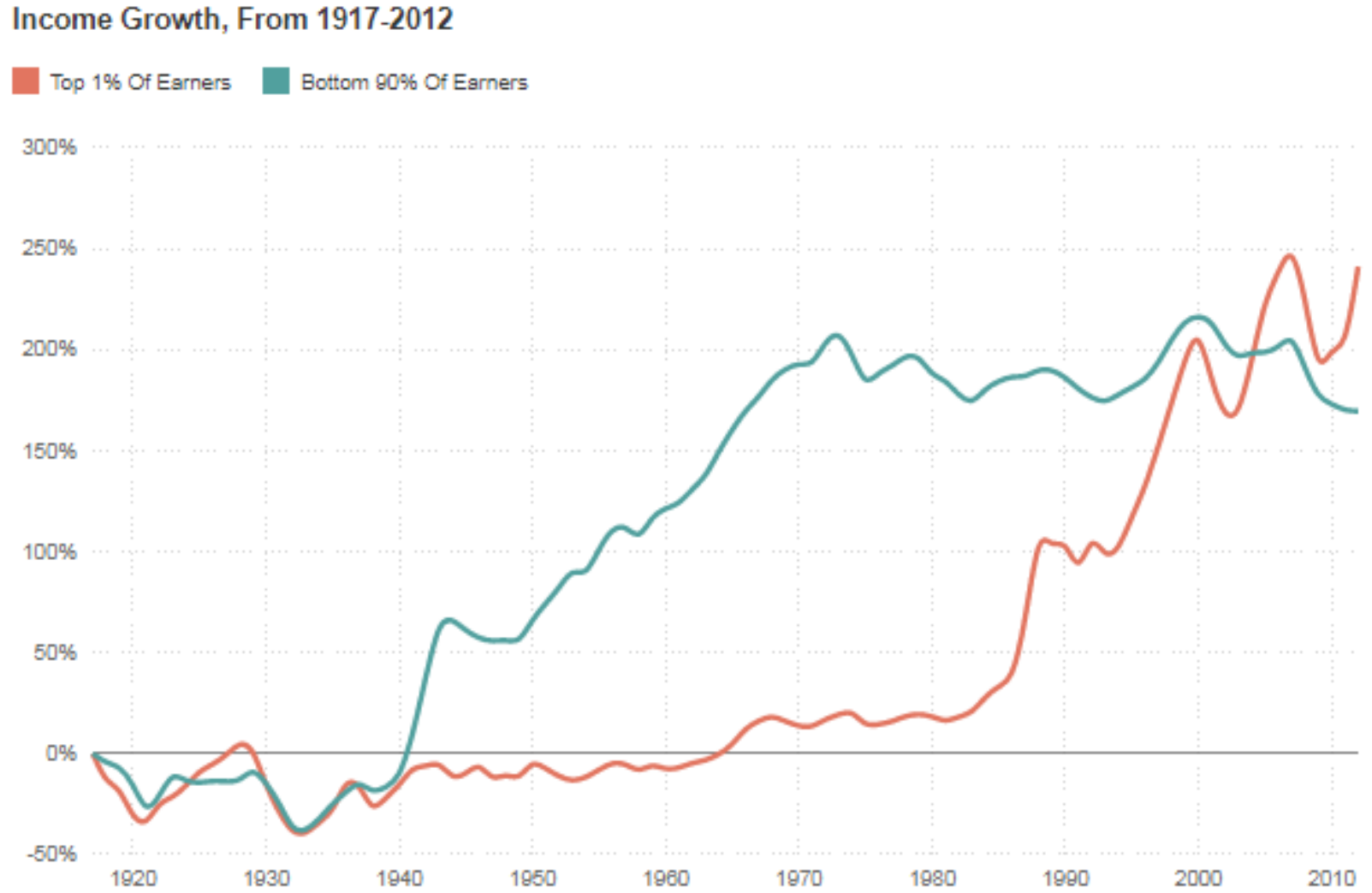
- Angle of analysis
  - Relevance: audience, context, message
  - Sufficiency: number of angles
- Framing
  - Reducing clutter
- Focus
  - Reducing noise

Visualisation Workflow - Failed Visualisations

- Useless solution
  - Failed to focus on relevant content
  - Not deep enough
  - Complex subject oversimplified
  - Not fit for setting
- Obstructive solution
  - Visually inaccessible
  - Misjudge format
  - Too many functions
- Not understandable
  - Too complex
  - Complex chart type
  - Absent annotations

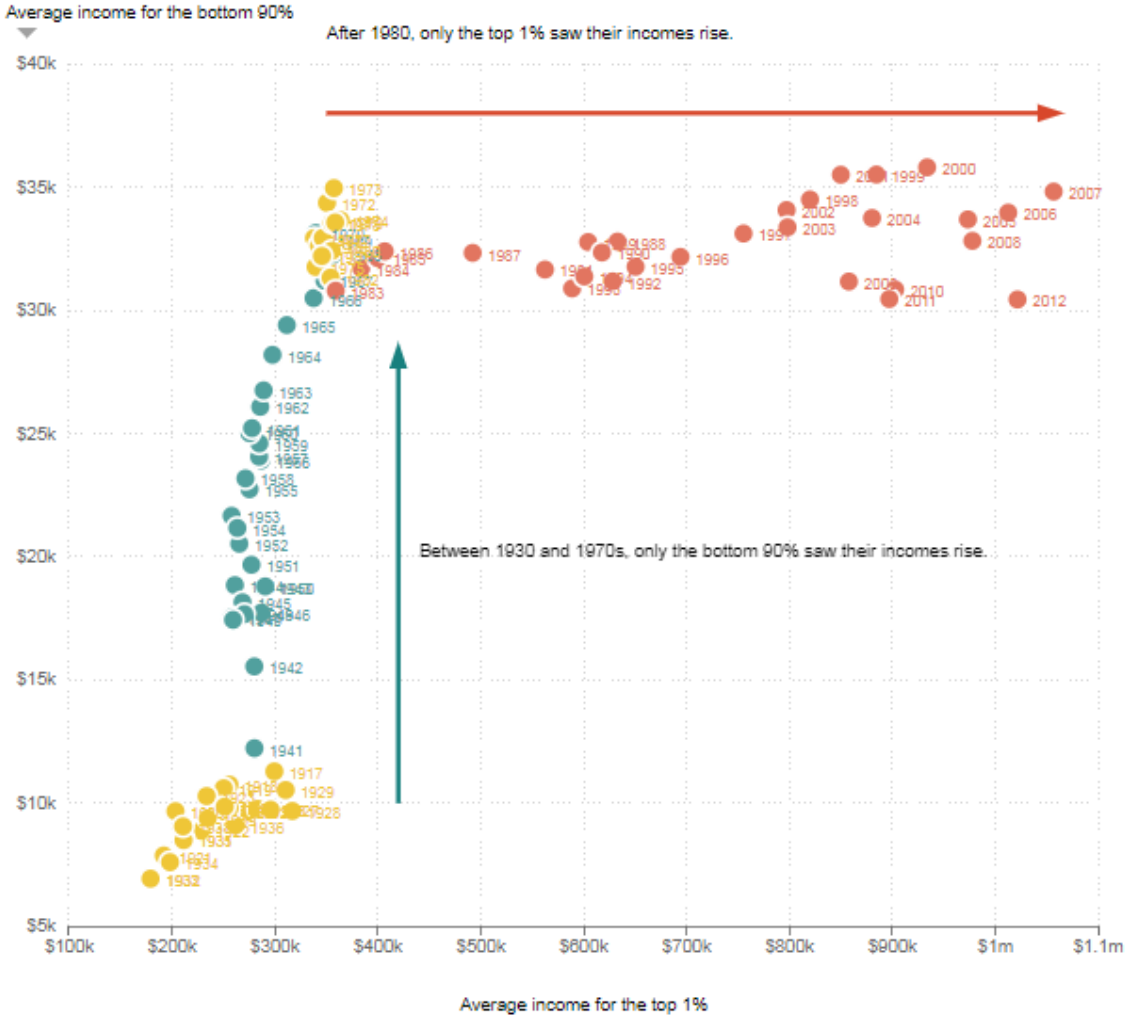
Data Visualisation: A Handbook for Data Driven Design. Andy Kirk

Case Study



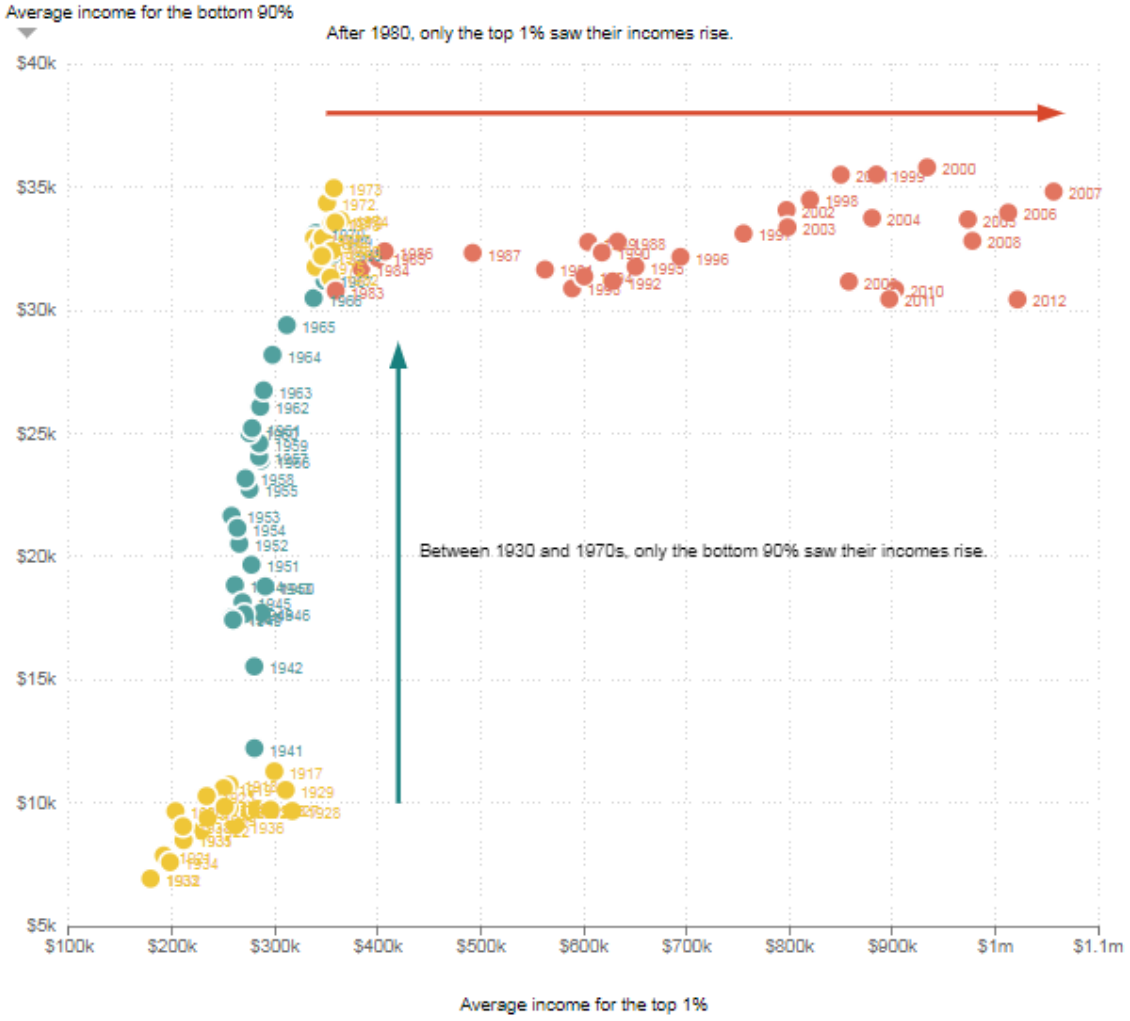
https://www.npr.org/sections/money/2015/02/11/384988128/the-fall-and-rise-of-u-s-inequality-in-2-graphs

Case Study



https://www.npr.org/sections/money/2015/02/11/384988128/the-fall-and-rise-of-u-s-inequality-in-2-graphs

Case Study



- Angle
  - Relationship between 2 measures and how it has changed over time
- Framing
  - USA
  - Time frame
- Focus
  - Colour showing two trends + annotations
  - Time slider

https://www.npr.org/sections/money/2015/02/11/384988128/the-fall-and-rise-of-u-s-inequality-in-2-graphs

CORRELATION & CAUSATION

Statistics & Relationships

- Statistics is about finding relationships in data
  - What are the similarities between groups?
  - Do they behave similarly?
  - Do they have opposite behaviours?

What Relationships To Look For?

Look for relationships between different variables

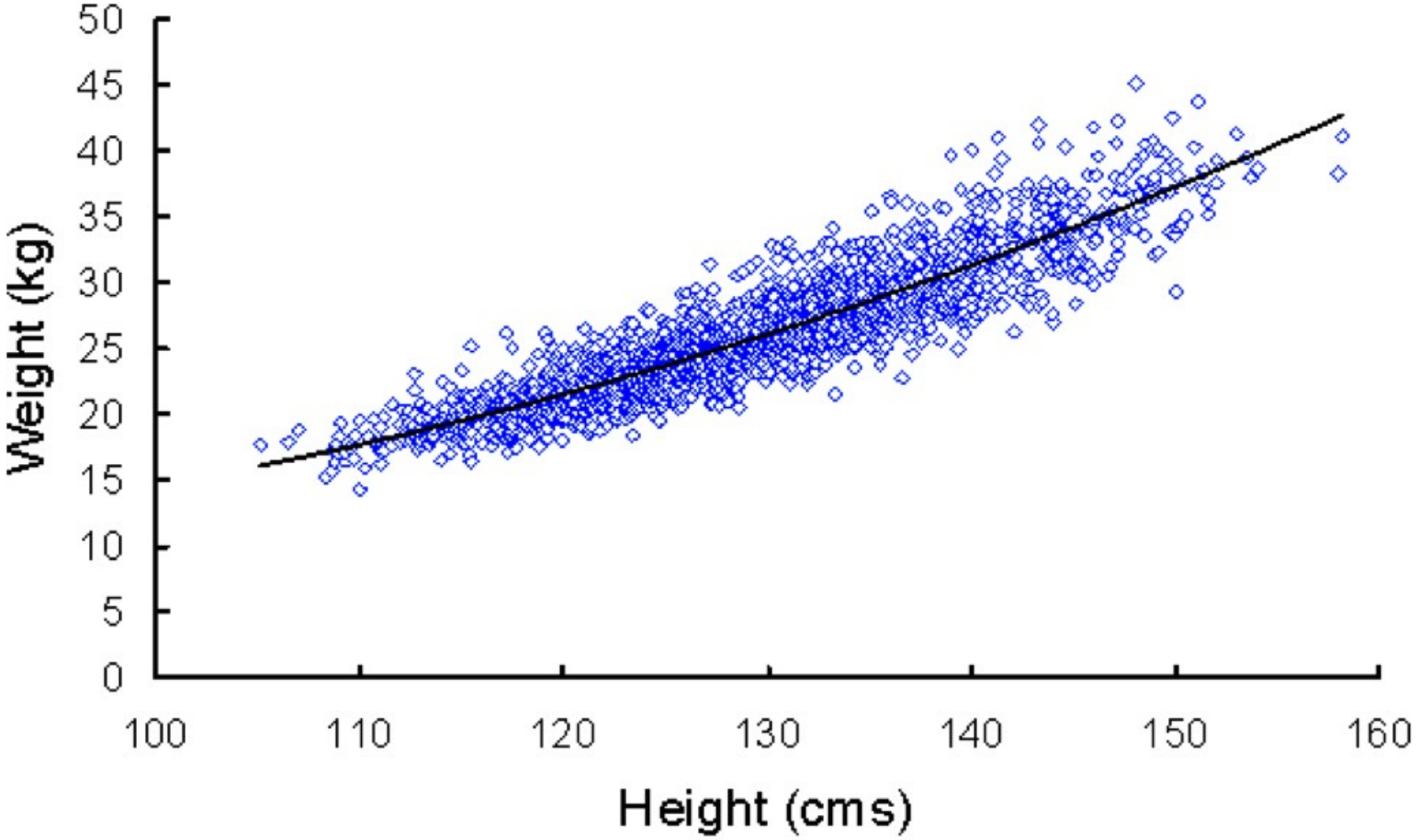
- As a variable goes up, does another variable go down?
- If so, is it a correlative or causal relationship?
  - You can show correlation relatively easily, which can lead to a deeper more exploratory analysis
  - A causal relationship is usually harder to prove quantitatively (which makes it even less likely you can prove it with a graphic)

What Relationships To Look For?

- Take a step back to look at the big picture - the distribution of your data
  - Is it spaced out or is it clustered in between?
  - Such comparisons can lead to stories about citizens of a country or how you compare to those around you.

- Compare multiple distributions for an wider view of your data
  - How has the makeup of a population changed over time?
  - How has it stayed the same?

Height & Weight



http://child-development.org/publishingimages/graphheightweight

Correlation & Causation

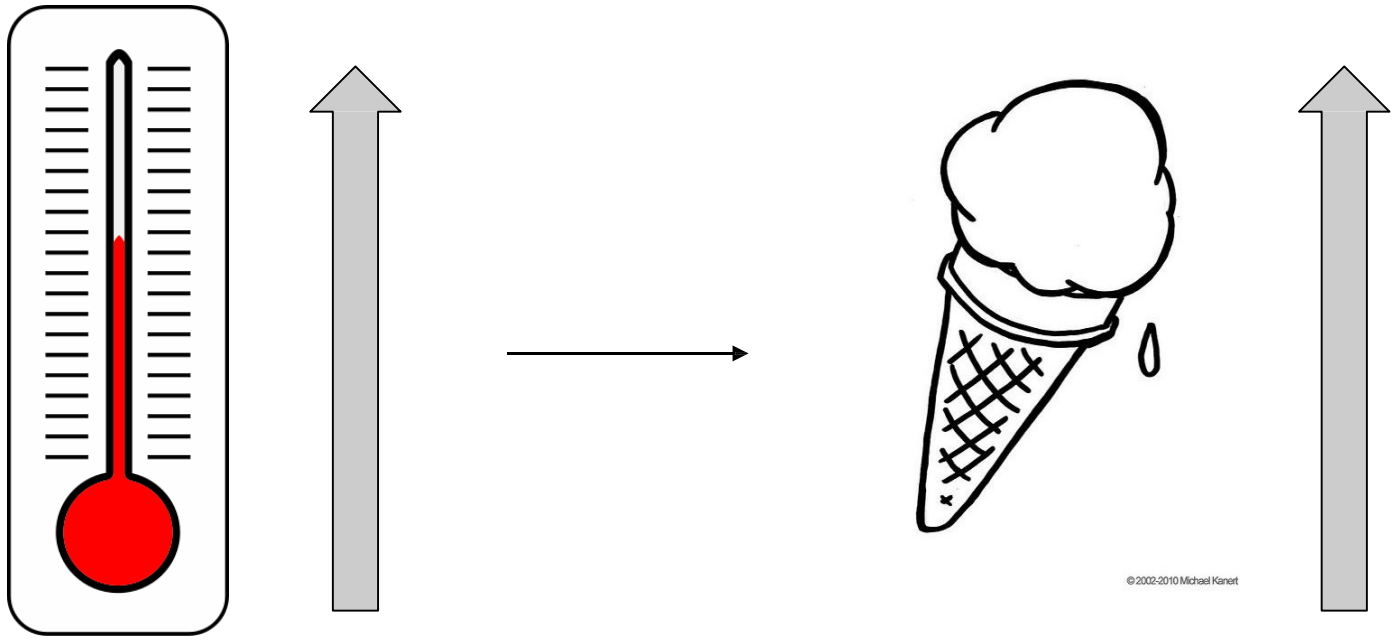
Correlation

- “A statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables.”

- Causation
- “Indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events. This is also referred to as cause and effect.”

Correlation & Causation

Correlation: one variable tends to change a certain way as another variable changes

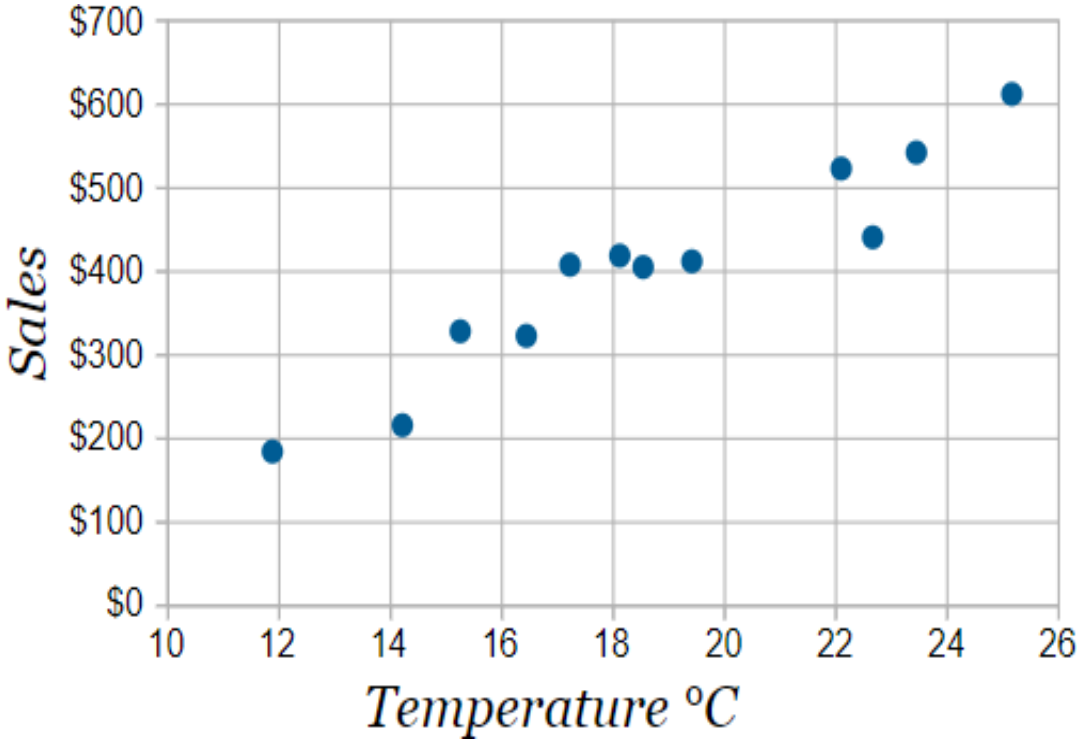


Independent variable

Dependent variable

Correlation & Causation

Correlation: one variable tends to change a certain way as another variable changes



Correlation & Causation

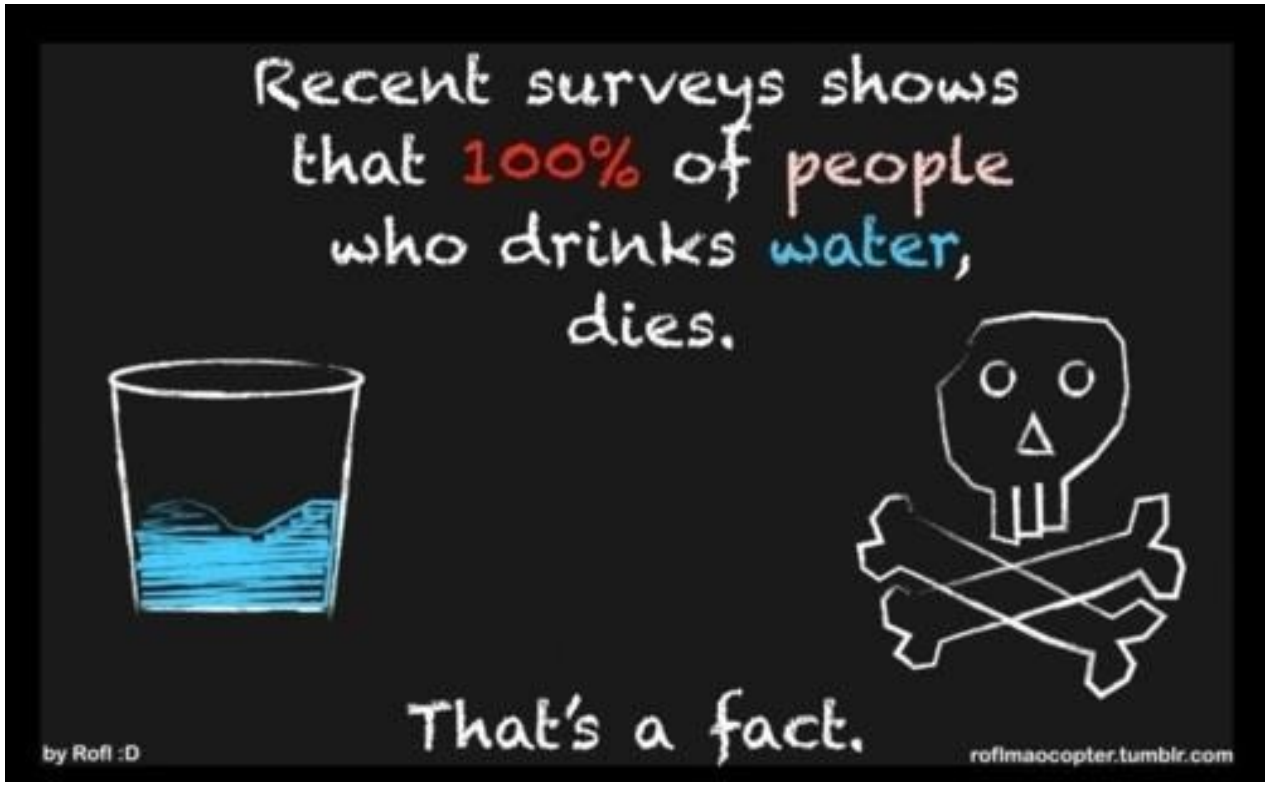
- Causation: one event is the result of the occurrence of the other event

- Smoking is correlated with alcoholism
  - Does smoking cause alcoholism?

- Smoking is related to an increased risk of developing lung cancer

Correlation & Causation

- Just because two things are connected, it doesn't mean that one caused the other



http://blog.lib.umn.edu/meyer769/section16&17/2011/12/correlation-vs-causation-1.html

Correlation & Causation

- **Extraneous variables** are variables that may compete with the independent variable in explaining the outcome of a study

- A **confounding variable** is an extraneous variable that does indeed influence the dependent variable

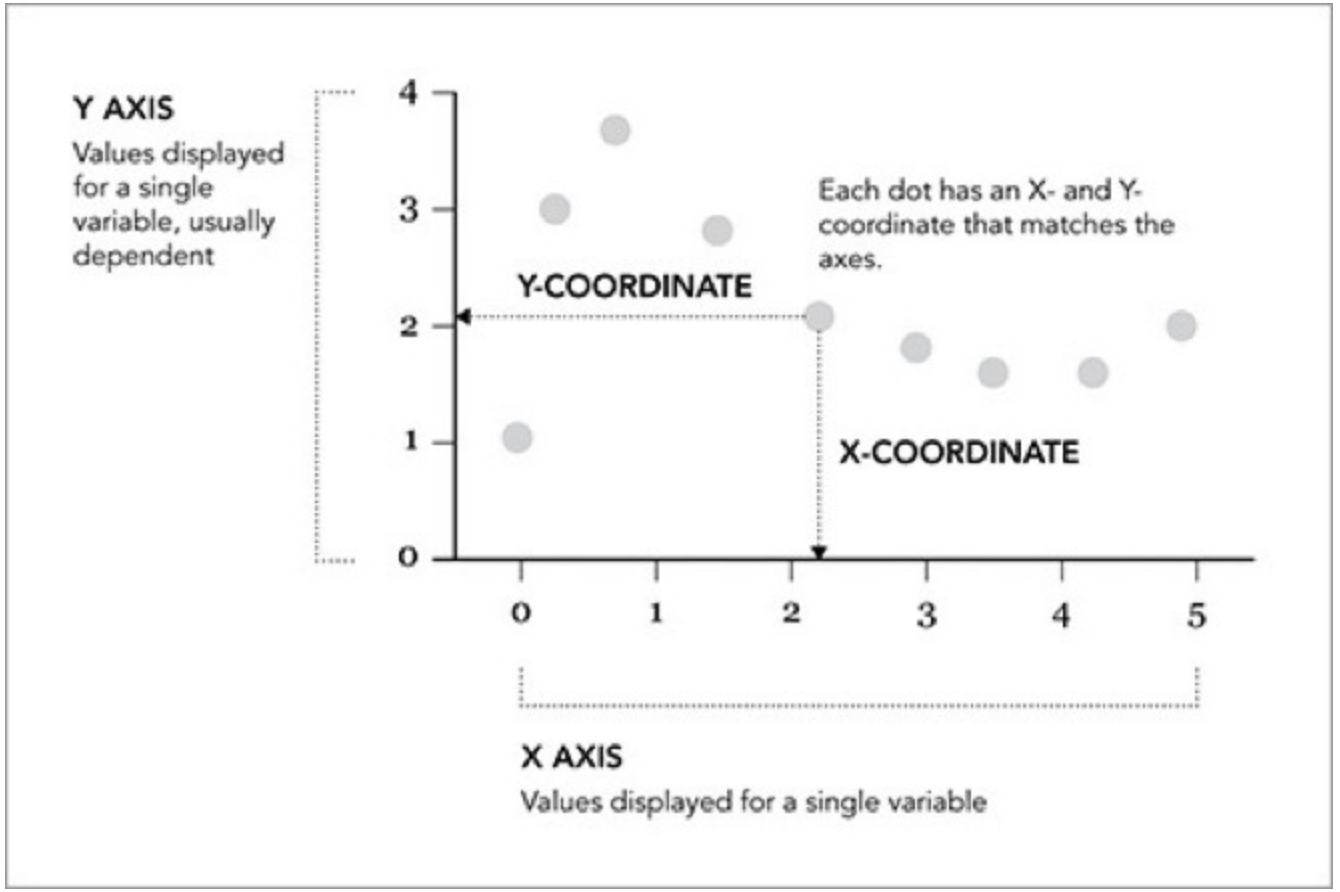
Finding Correlation

- It’s difficult to account for every outside, or confounding factor, which makes it difficult to prove **causation**

- You can, however, easily find and see correlation and a **scatter plot** is our key tool for visualising it

SCATTER PLOTS

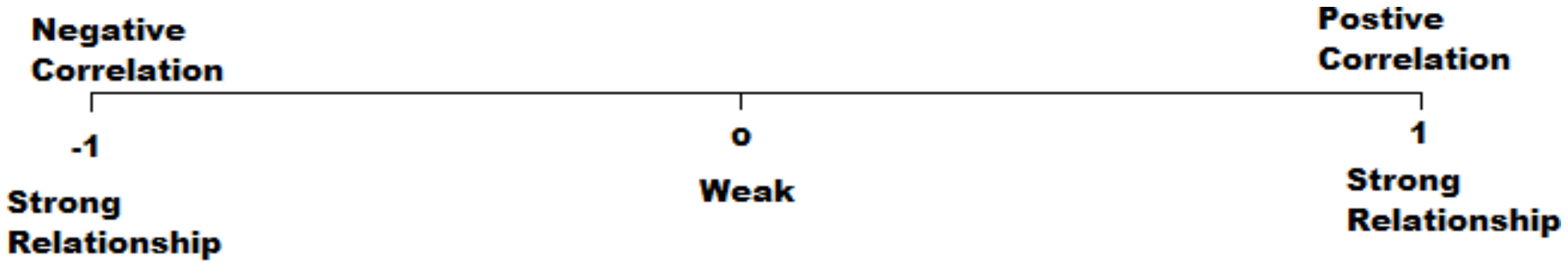
Simple Scatter Plot

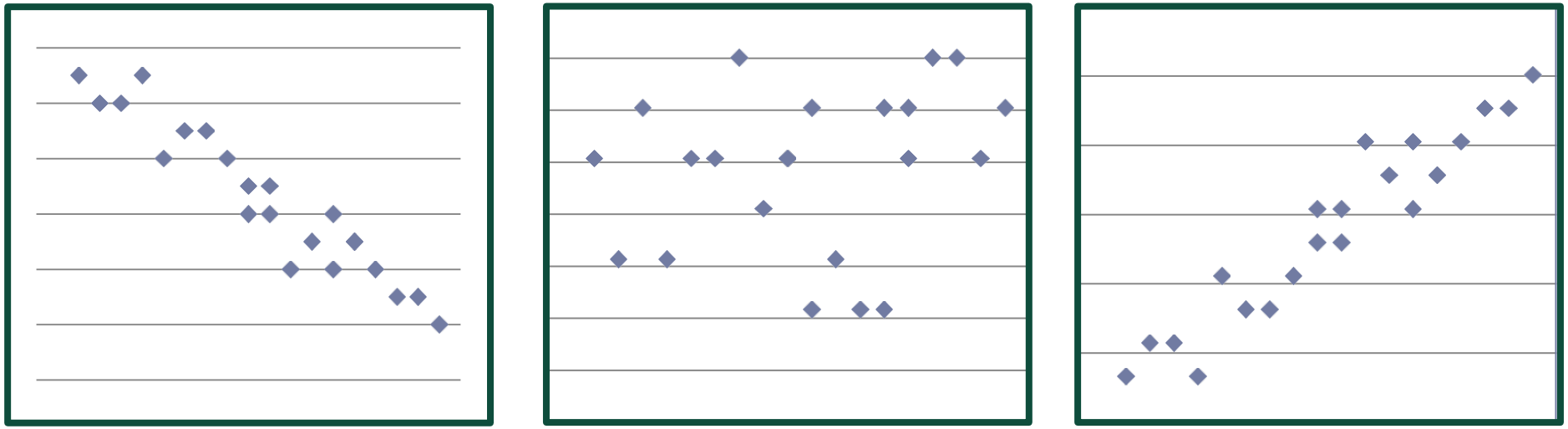


Displays relationship between two quantitative measures for different categories

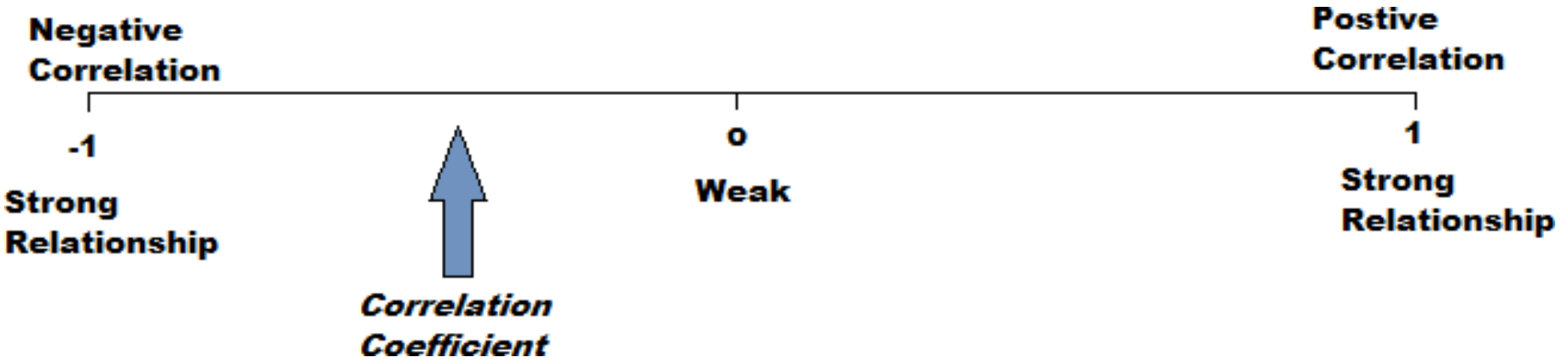
Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics. Nathan Yau. 2011

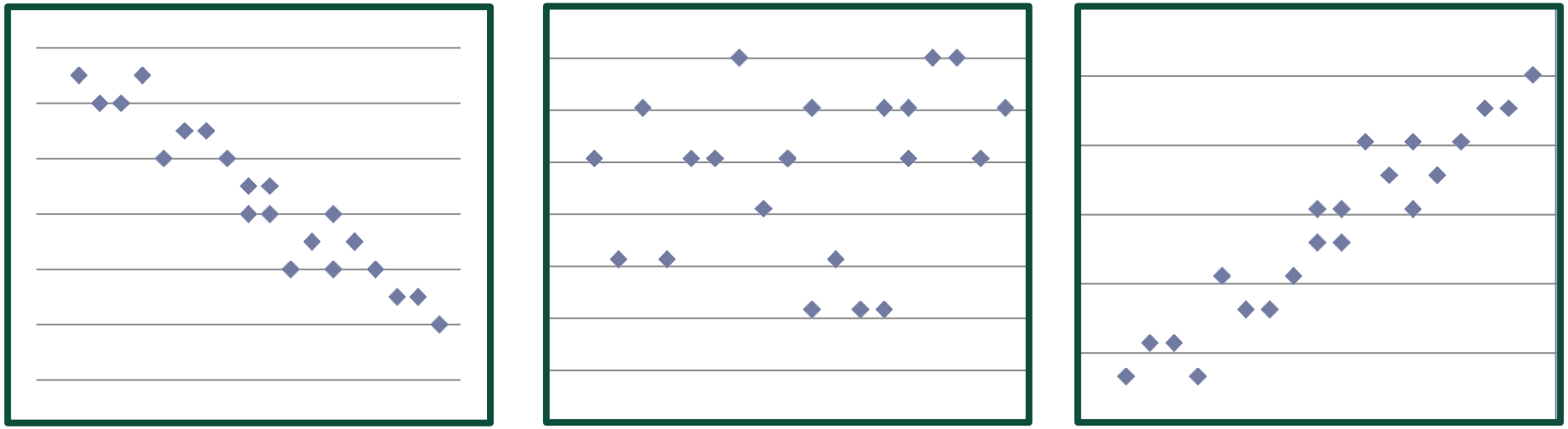
Simple Scatter Plot





Simple Scatter Plot



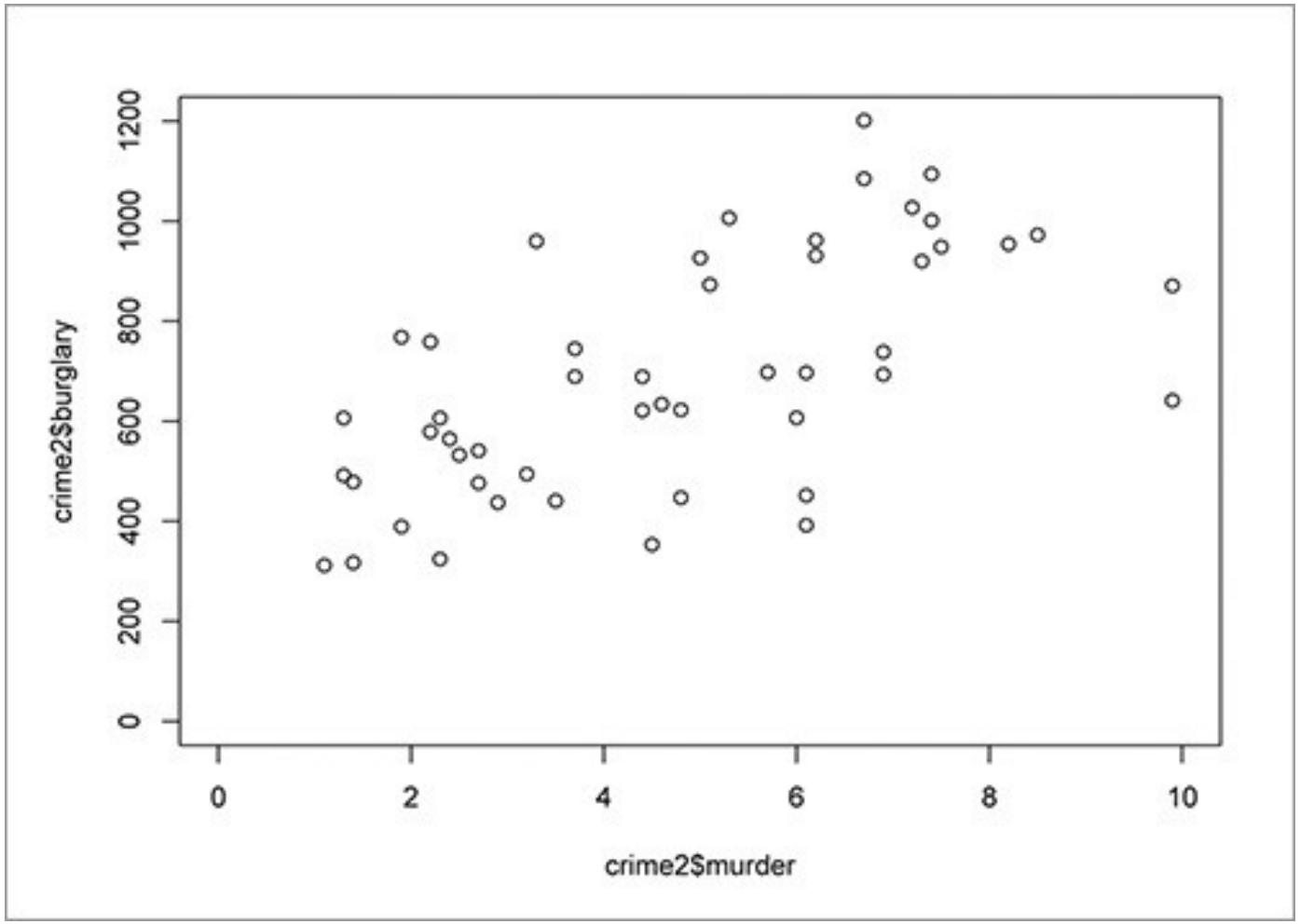


Simple Scatter Plot

- Scatter Plots do not work well if one or both measures have limited variation in value (occlusion problems)

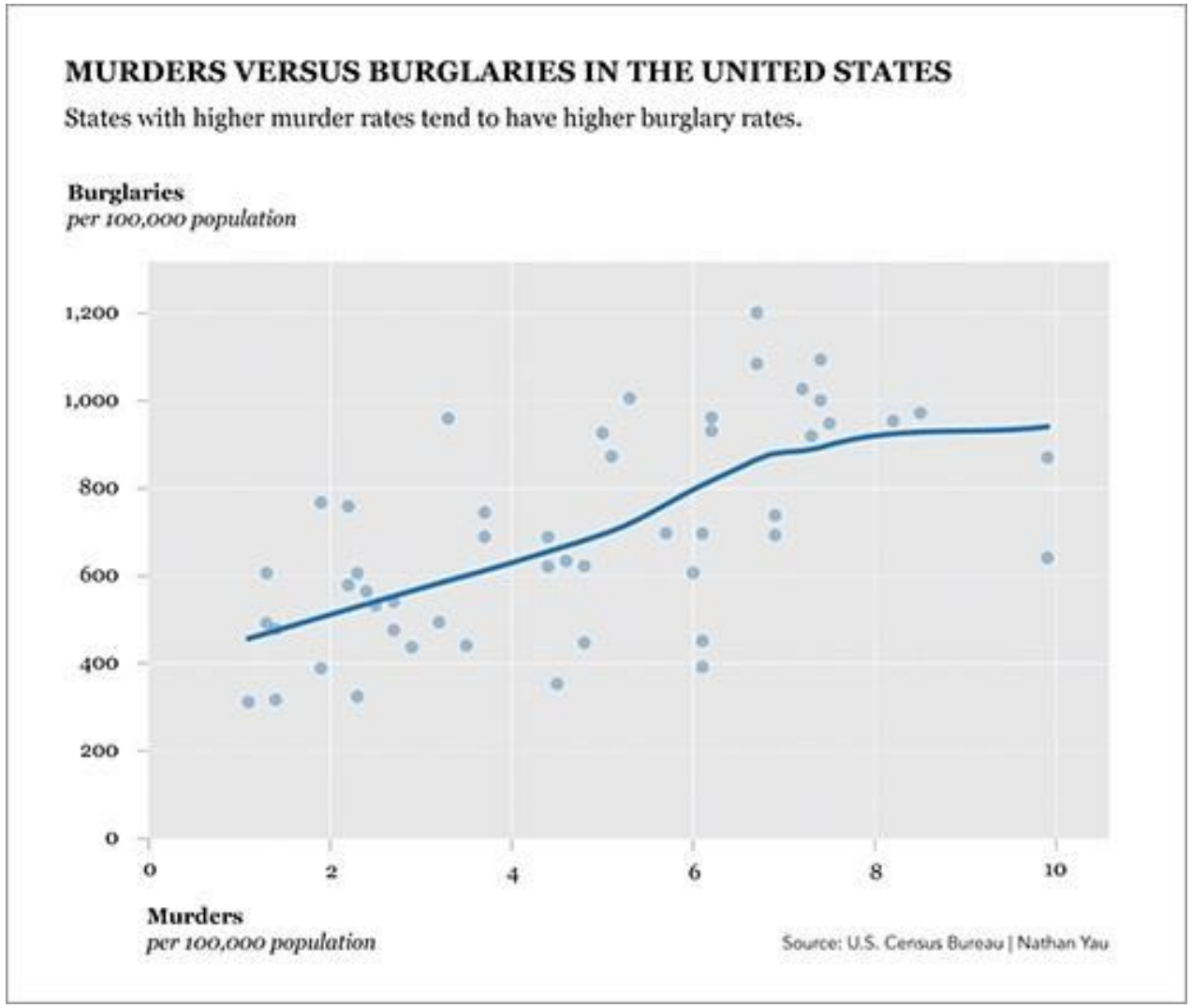
- Composition
  - X- independent variable
  - Y- dependent variable
  - 1:1 aspect ratio
  - No need to start at 0

Example: US Crime Rates



Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics. Nathan Yau. 2011

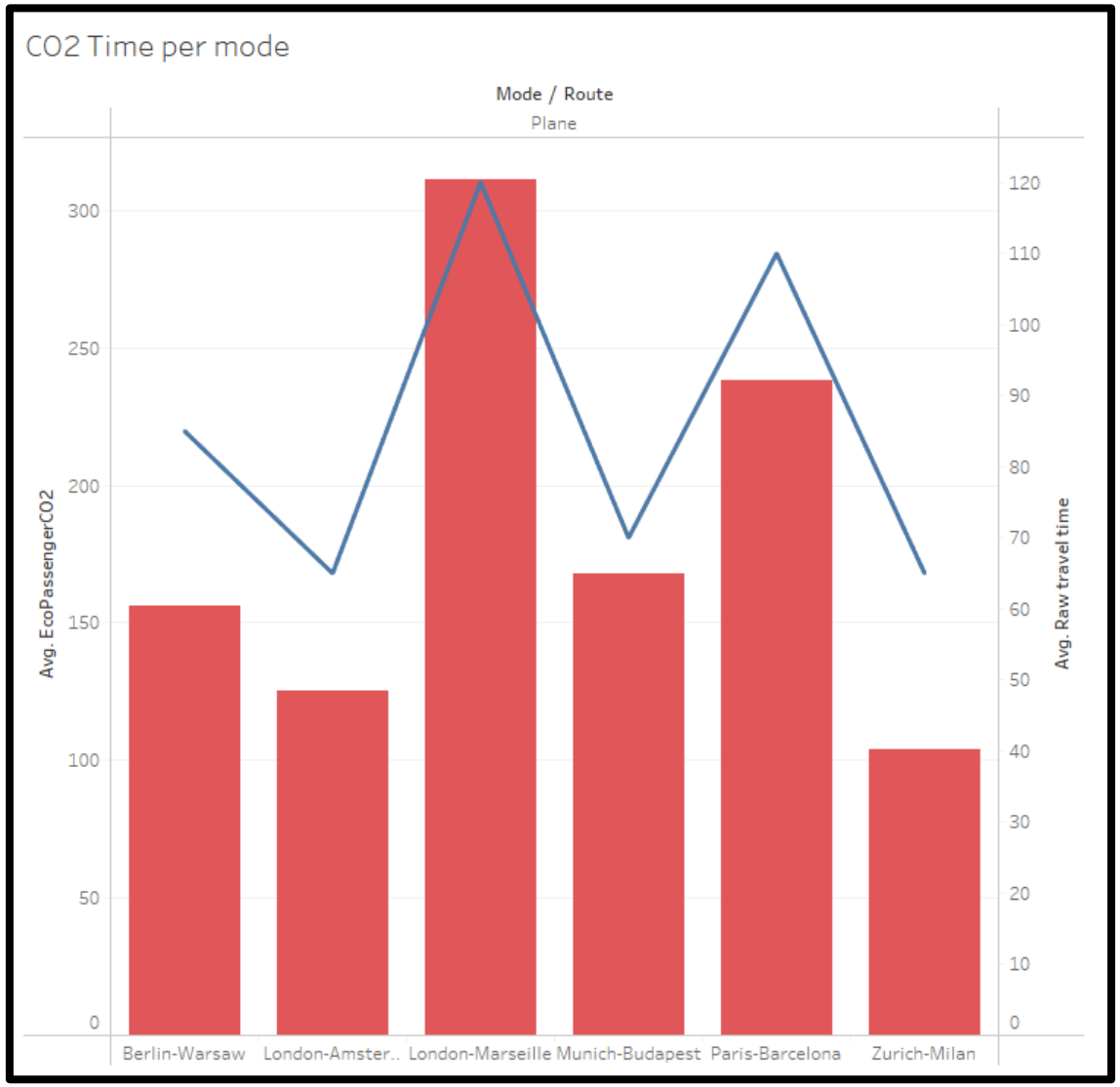
Example: US Crime Rates



Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics. Nathan Yau. 2011

LINE COLUMN CHARTS

Line Column Charts

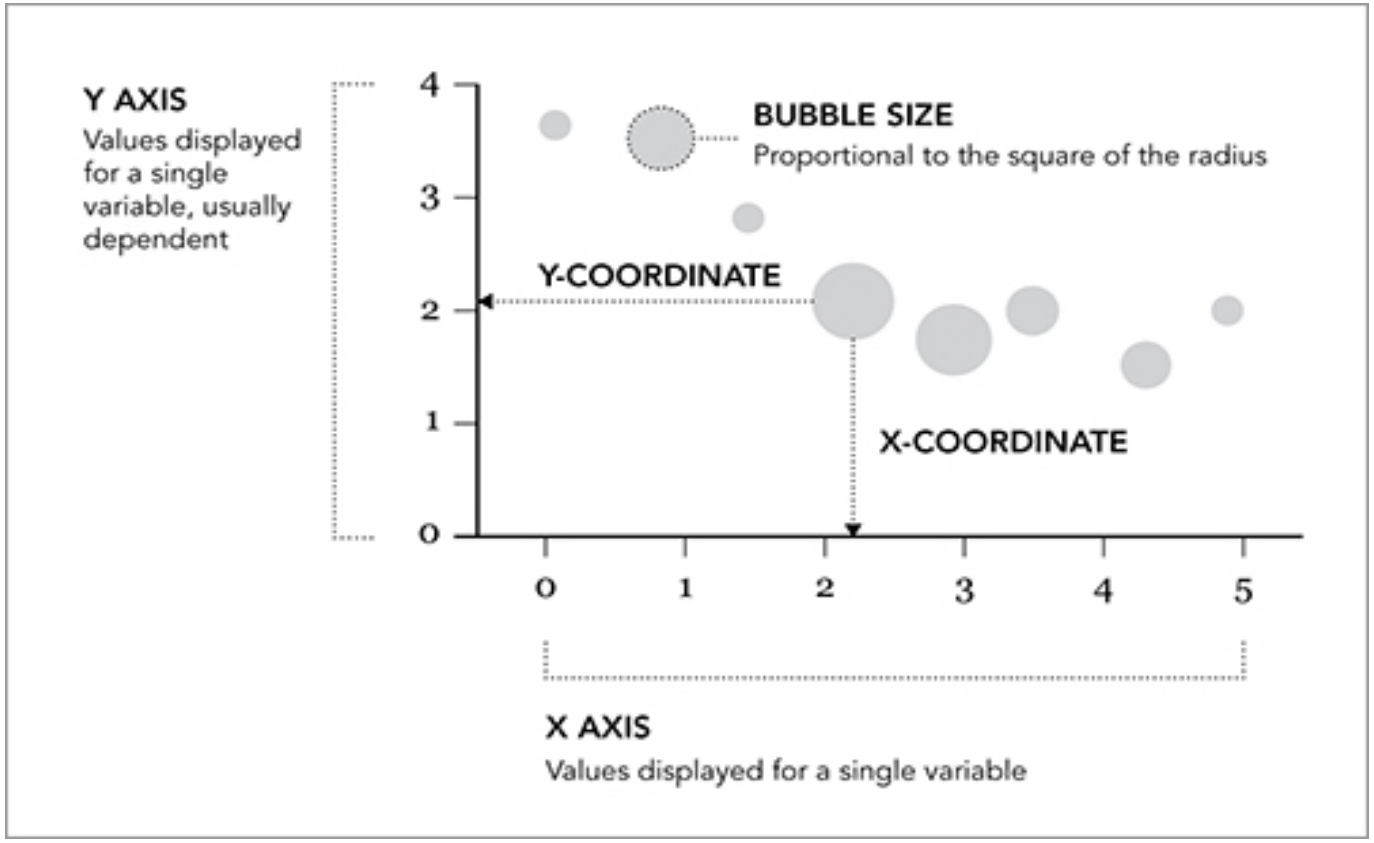


- Easily illustrates the relationships between two variables with different magnitudes and scales of measurement
- Note secondary axis

BUBBLE PLOTS

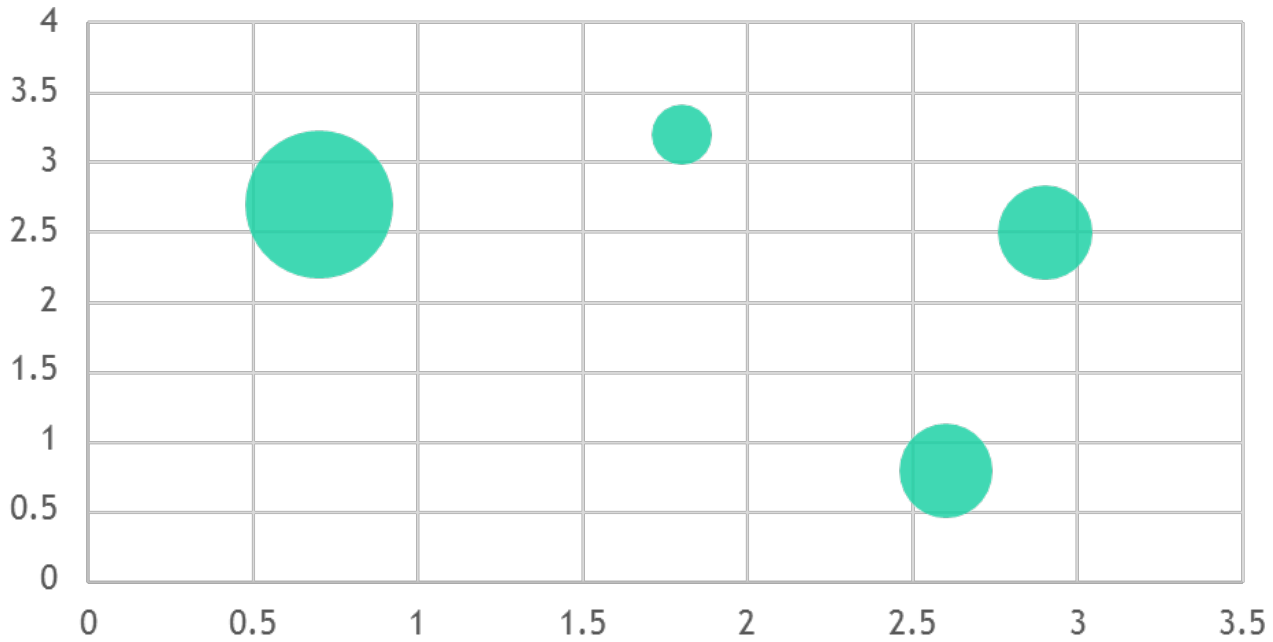
Bubble plots

- A bubble plot can be defined as a 3D scatterplot
  - The value of an additional variable is represented through the size of the dots.



Bubble plots

- A bubble plot can be defined as a 3D scatterplot
  - The value of an additional variable is represented through the size of the dots.



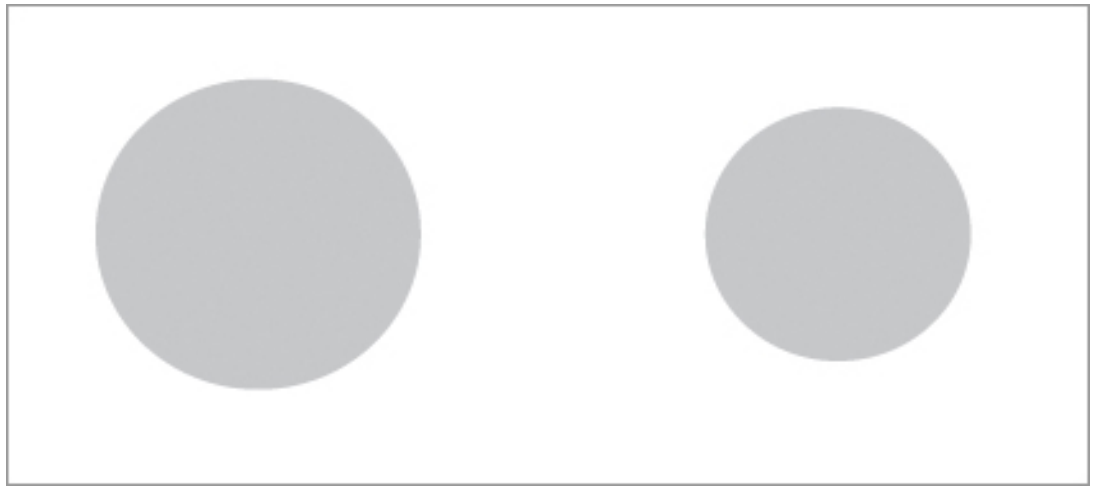
Bubble plots - Composition

- Too many bubbles make the chart hard to read
- X- independent variable, Y- dependent variable
- 1:1 aspect ratio
- No need to start at 0
- Add a legend to make possible the link between the size and the value
- The **area** of the circles must be proportional to the **value**, not to the **radius**, to avoid exaggerate the variation in your data

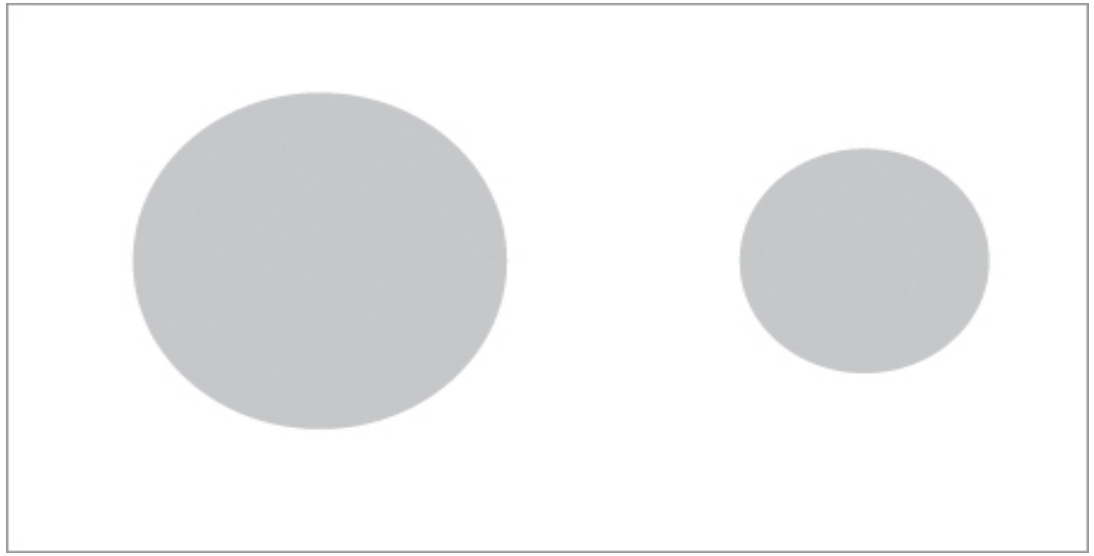
Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics. Nathan Yau. 2011

Bubble Plots

Sized by Area

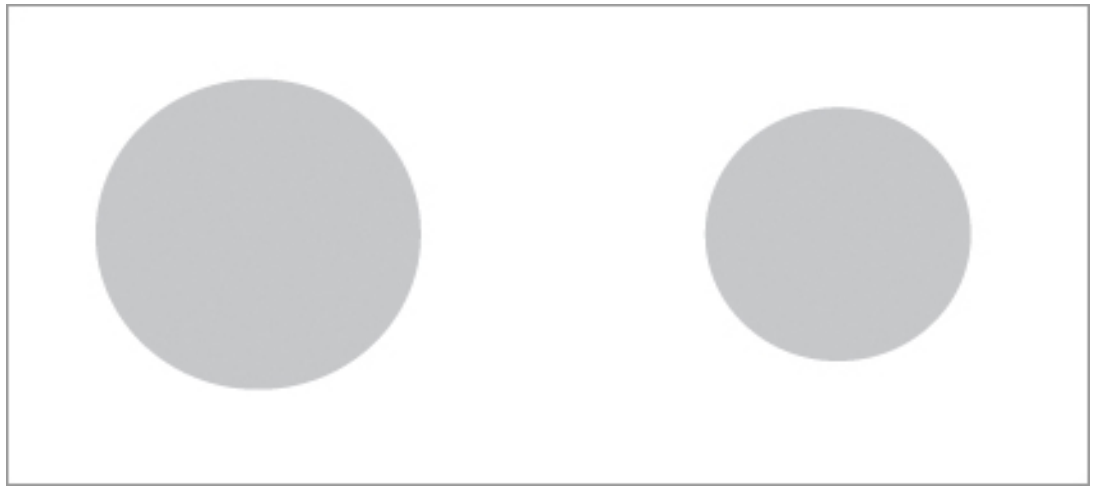


Sized by Radius

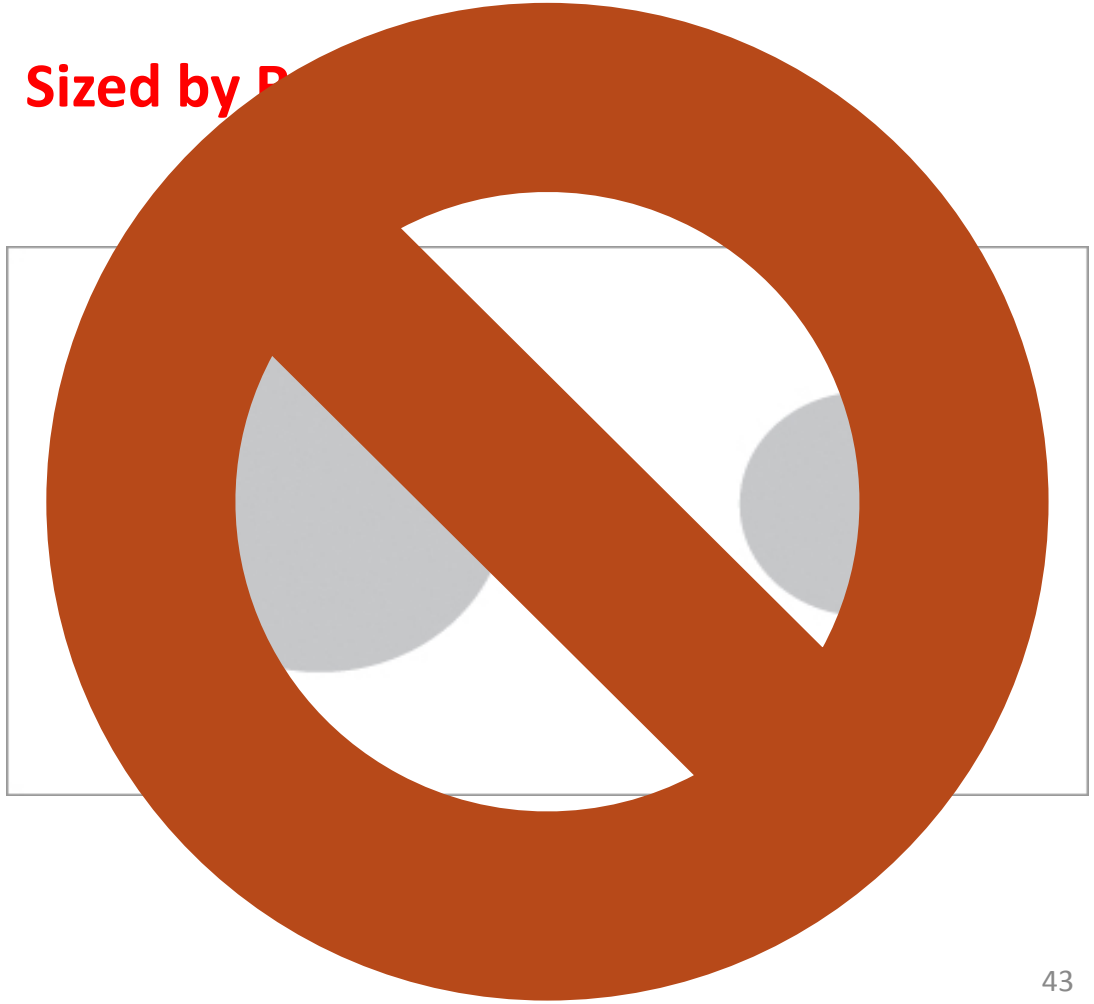


Bubble Plots

Sized by Area

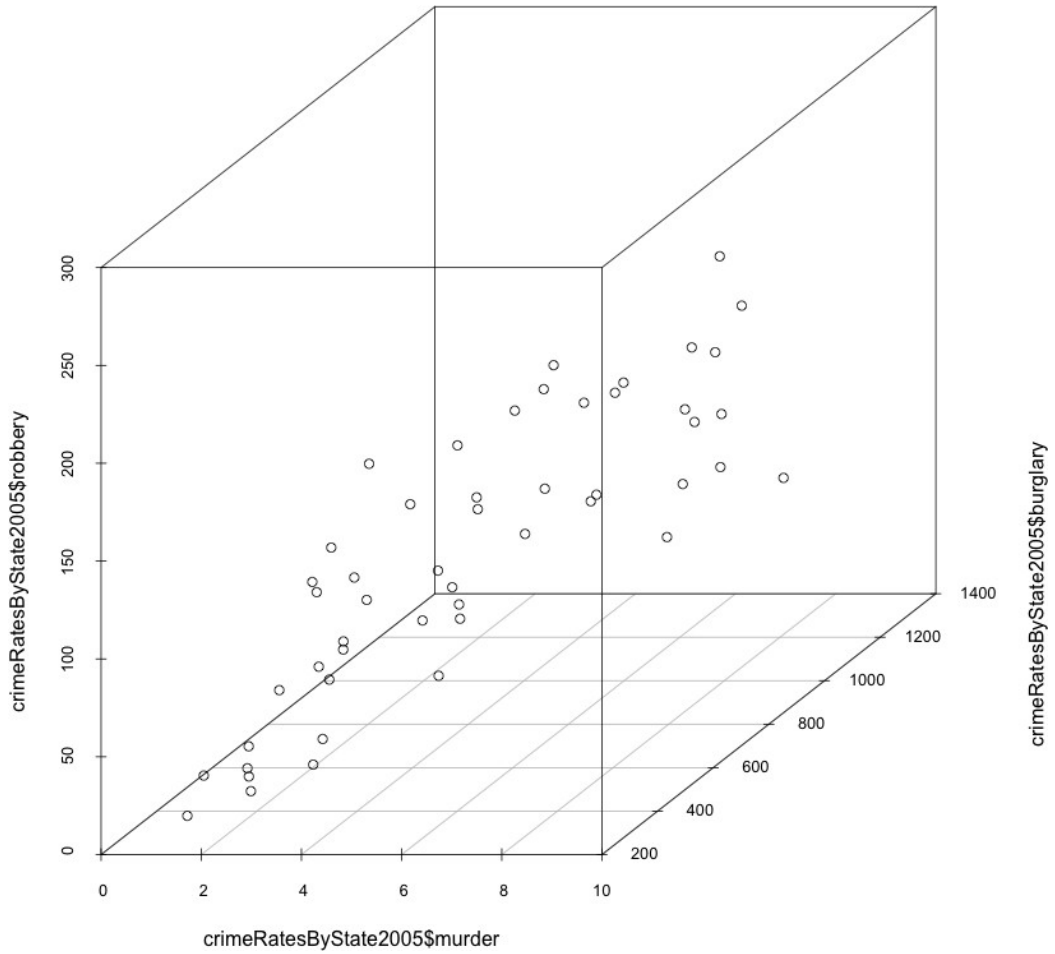


Sized by R



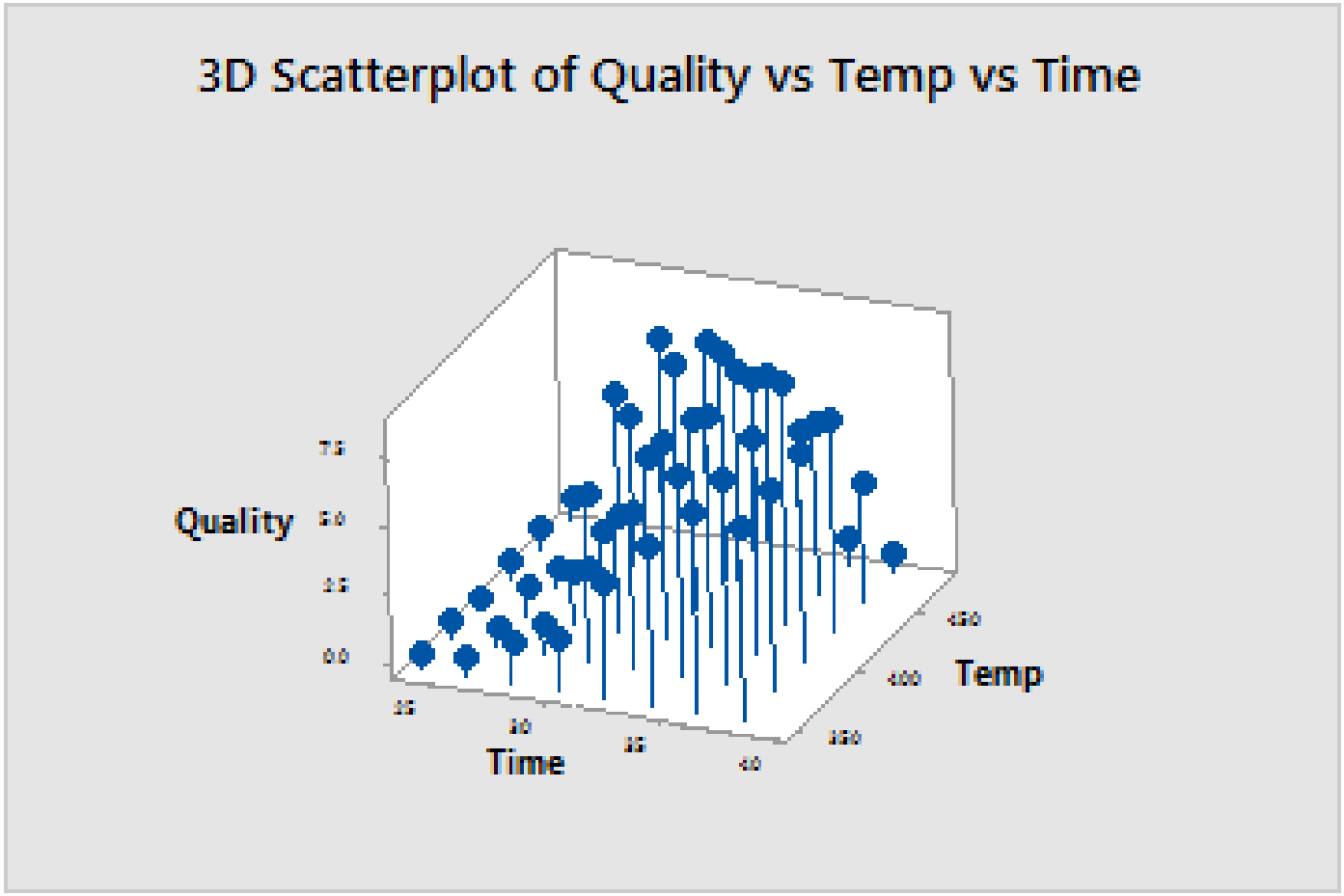
3D SCATTER PLOTS

3D scatter Plots



Adapted From Quick R: http://www.statmethods.net/graphs/scatterplot.html

3D scatter Plots



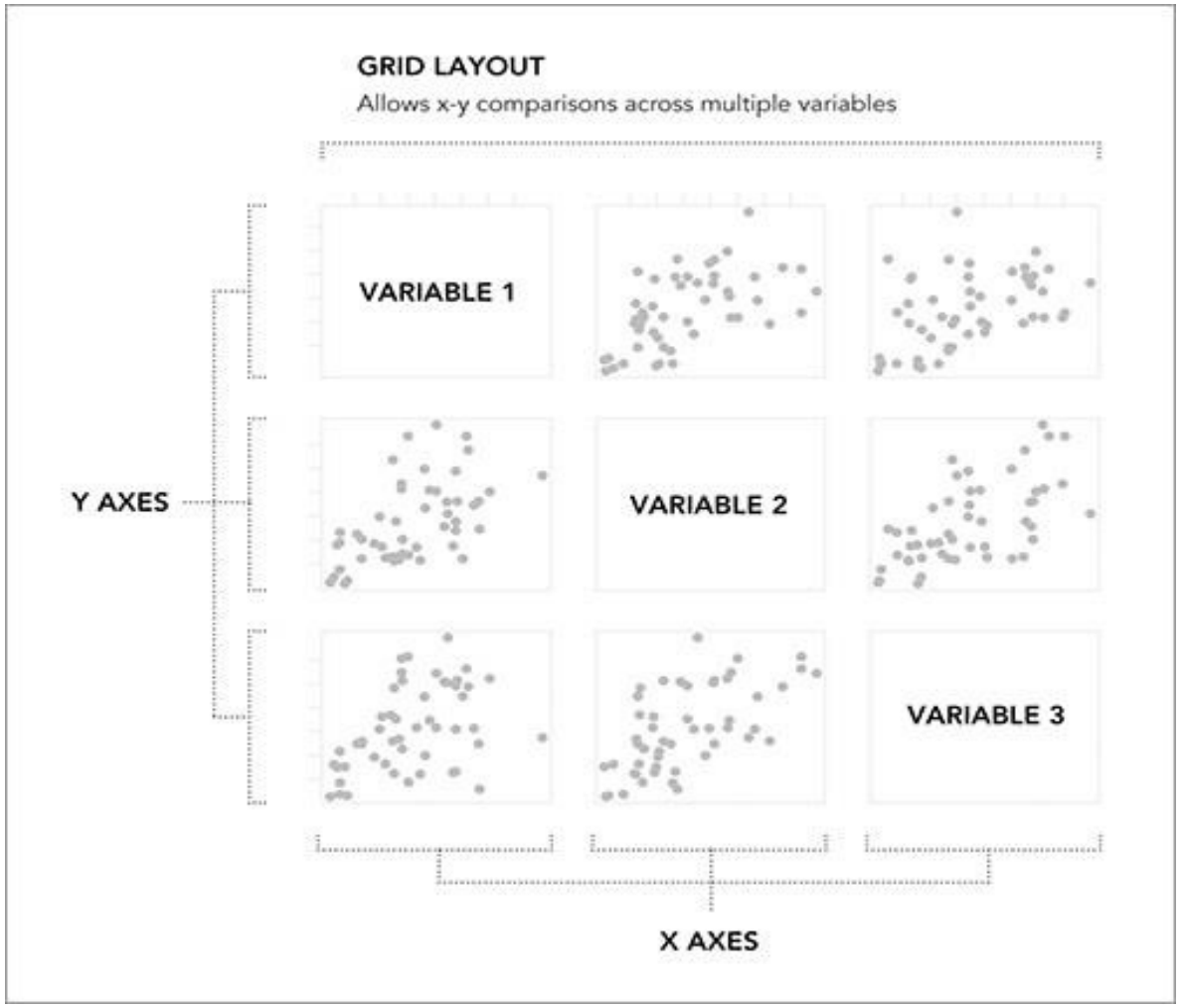
https://support.minitab.com/en-us/minitab/19/help-and-how-to/graphs/3d-scatterplot/before-you-start/example/

EXTRA DIMENSIONS

Exploring Even More Variables

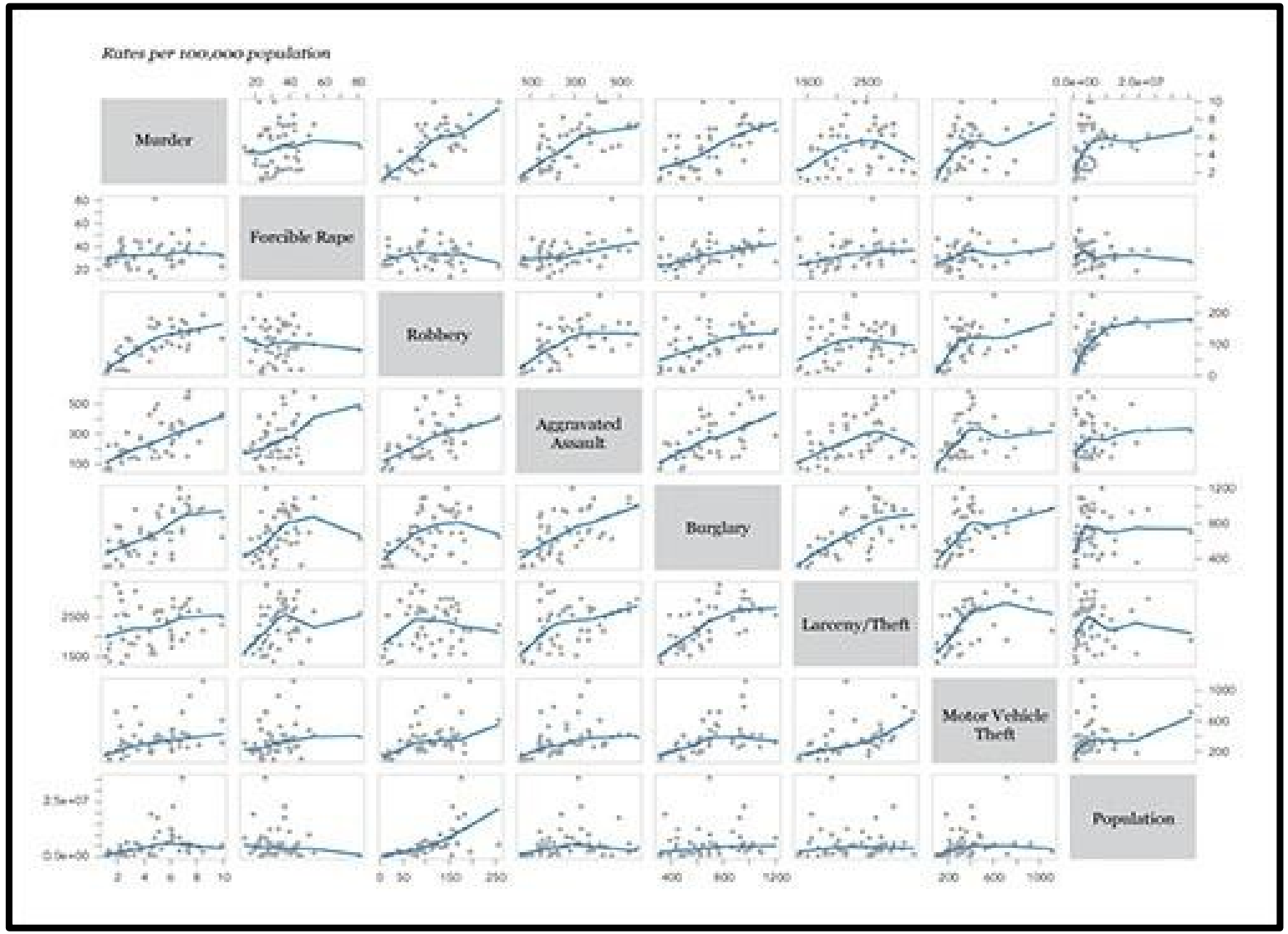
- You can plot every possible pair with a scatter plot matrix to compare all variables
- It’s usually a square grid with all variables on both the vertical and horizontal
- Each column represents a variable on the horizontal axis, and each row represents a variable on the vertical axis
- This provides all possible pairs

Scatter Plot Matrix



Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics. Nathan Yau. 2011

Example: US Crime Rates

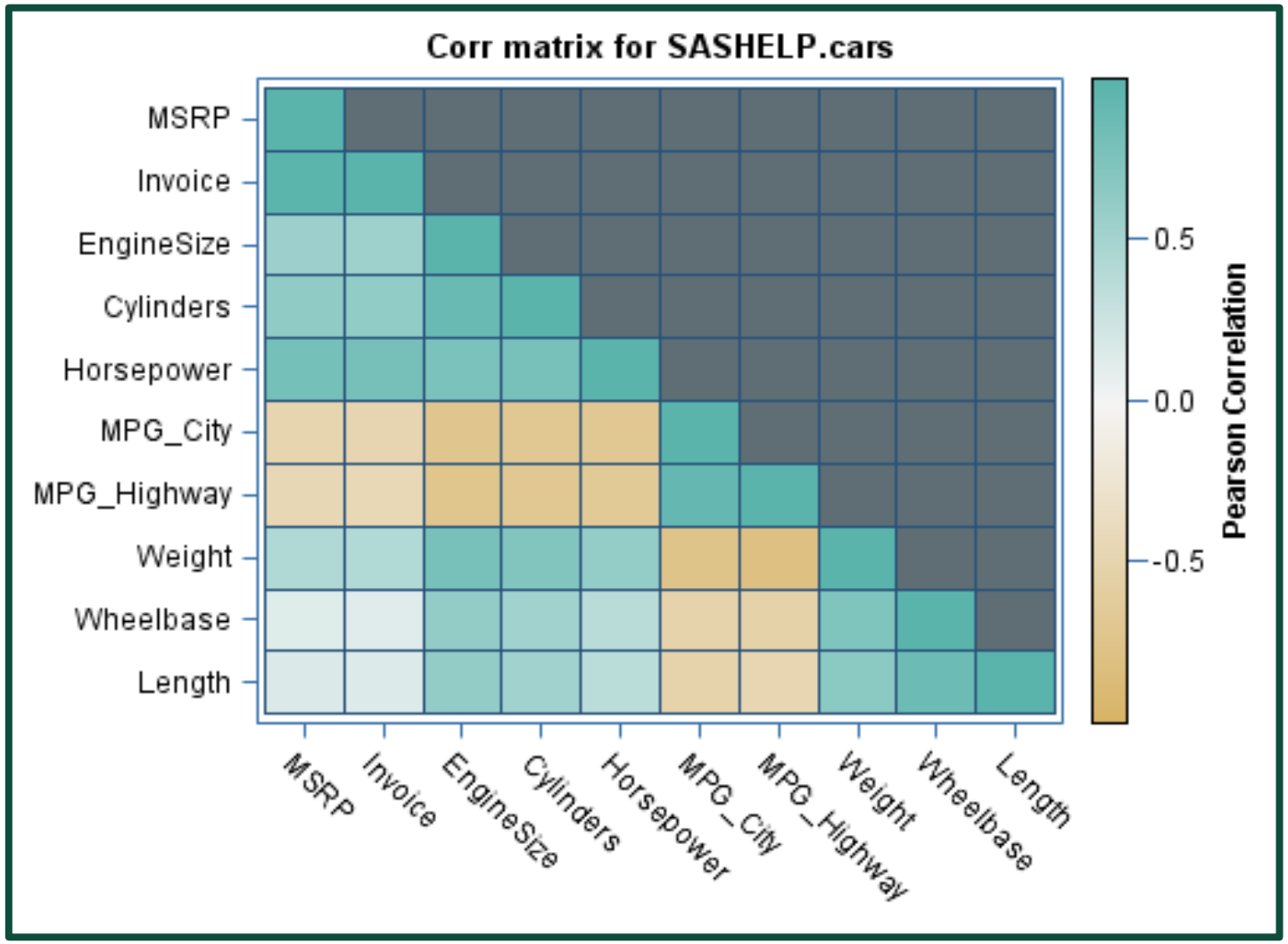


X Y HEAT MAPS

X Y HEAT MAPS

- A heat map displays quantitative values at the intersection between two categorical dimensions
- Two categorical axis with all possible values
- Each cell is colour coded to represent a quantitative value for each combination of category pairing

X Y HEAT MAPS



https://blogs.sas.com/content/sasdummy/files/2013/06/base_corrmatrix.png

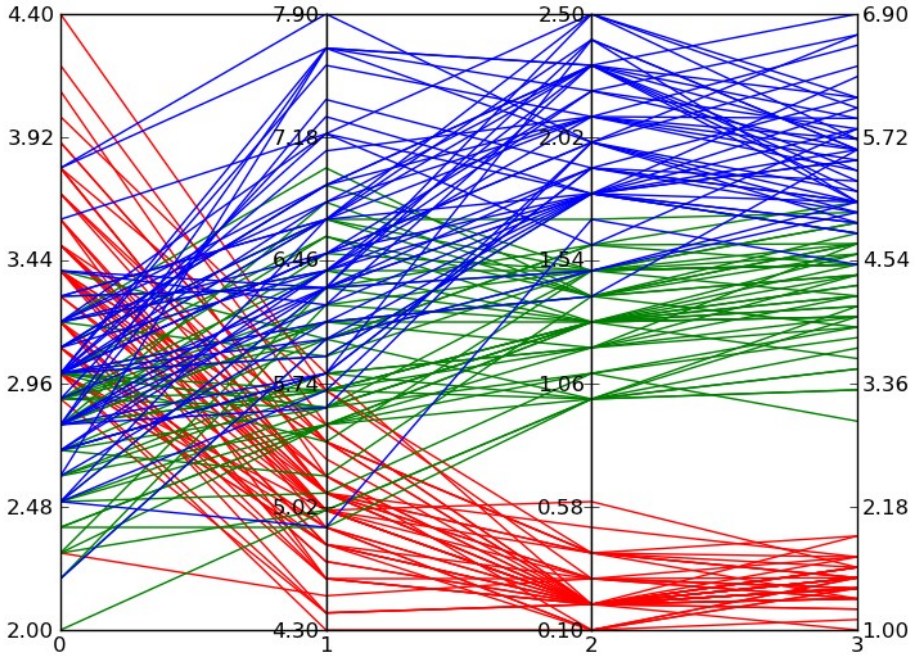
X Y HEAT MAPS

- Not easy to identify exact quantities represented by colours
- Order of magnitude information
  - Useful for finding patterns
  - Not good at showing fine differences in amounts
- Composition:
  - Logical sorting and sub-grouping can aid readability
  - Colour scale

PARALLEL COORDINATES

Parallel Coordinates

- Display of multiple quantitative measures for different categories in a single display
- Useful for exploratory analysis of multivariate data



https://i.stack.imgur.com/N1mpi.png

Parallel Coordinates

- Particularly useful when interactivity is added to the chart

- Composition:
  - The ordering of the variables has an effect on the patterns
  - Neighbouring measures should have a common scale and similar meaning
  - The more variables added the more difficult it will be to decipher

- No Tableau native chart

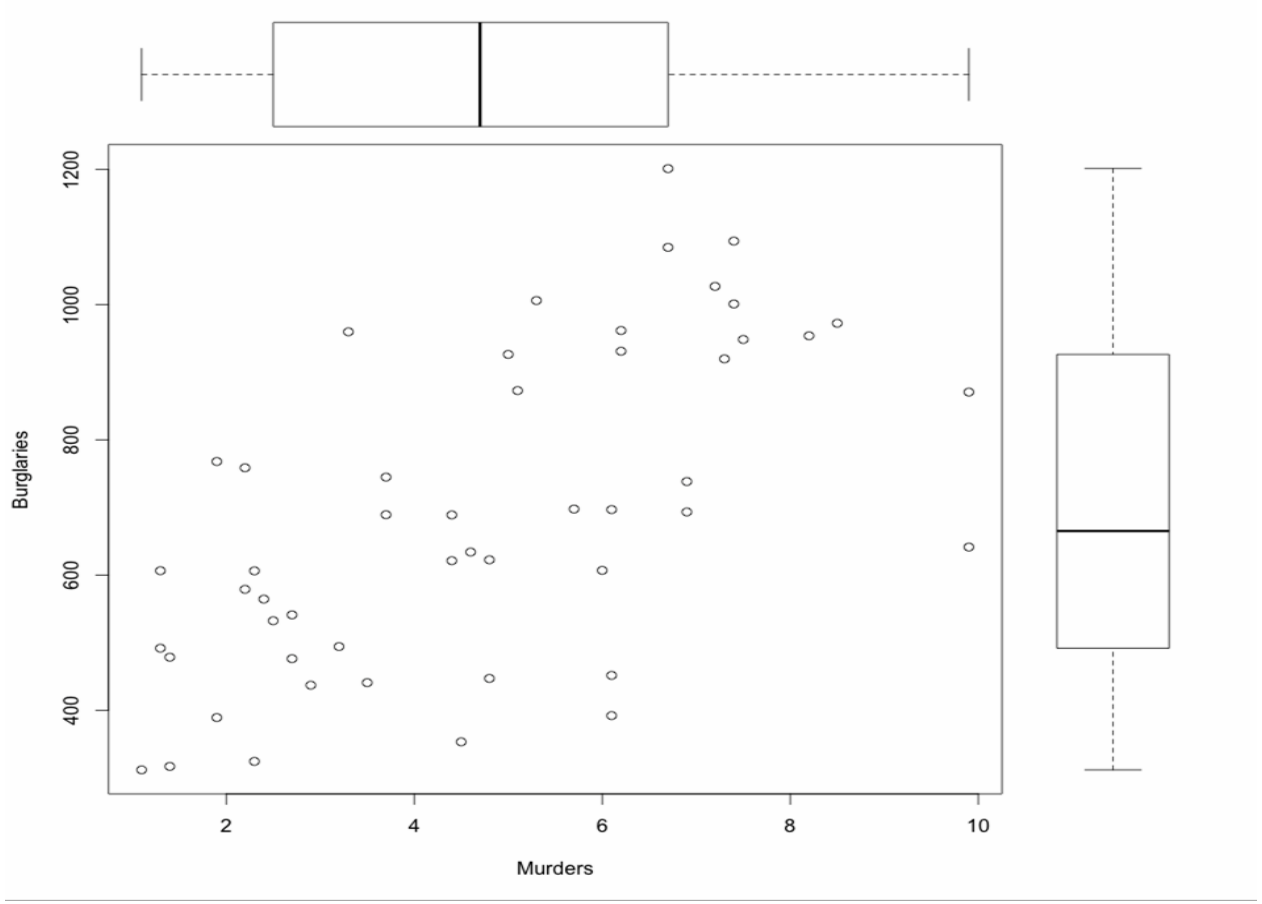
EXTENSIONS TO SCATTER PLOTS

Playing With Scatter Plots

- Scatter plots are our core tool for showing **relationships** or **correlations**

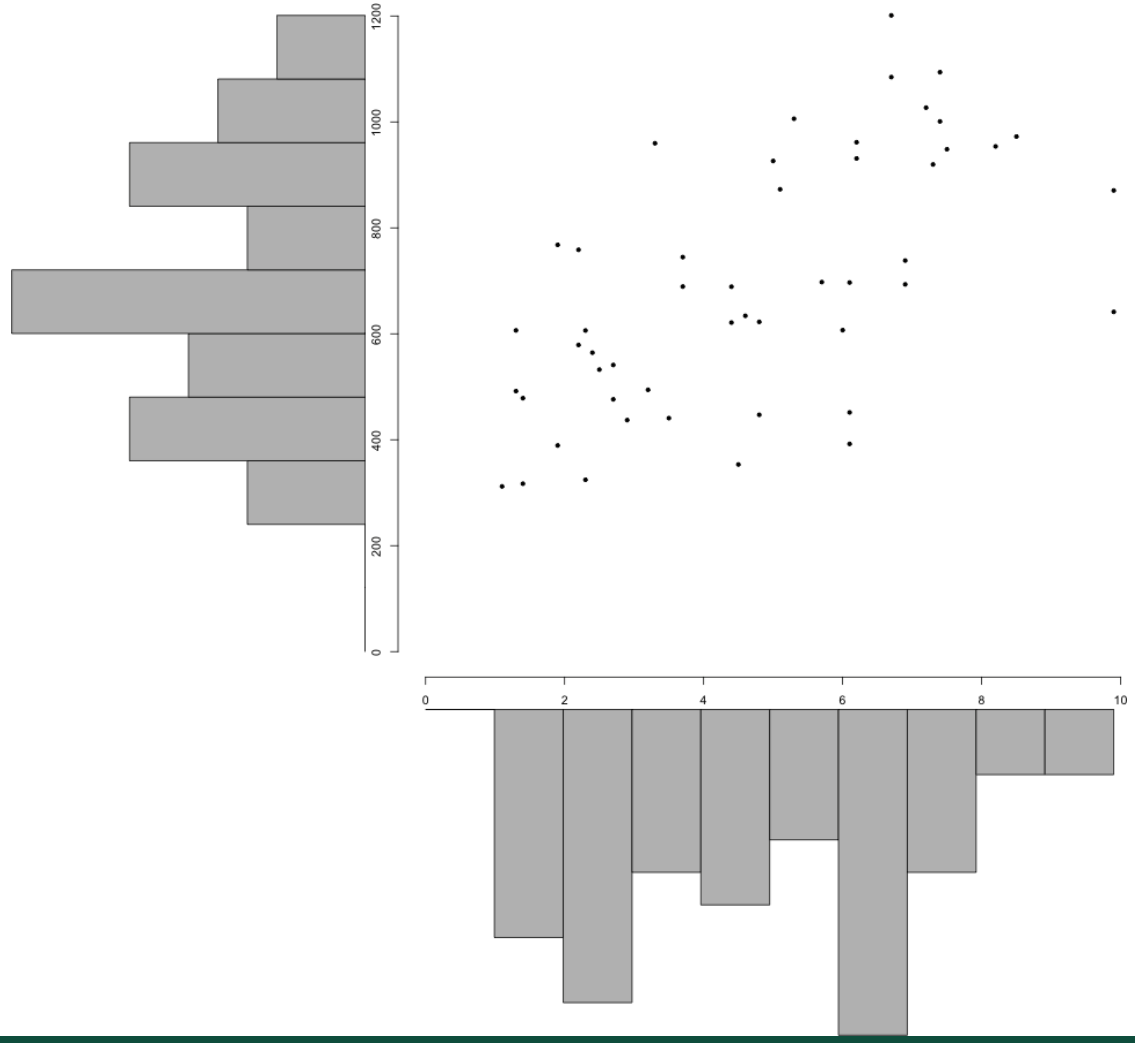
- We can augment scatter plots with other interesting things to show more information

Example: US Crime Rates



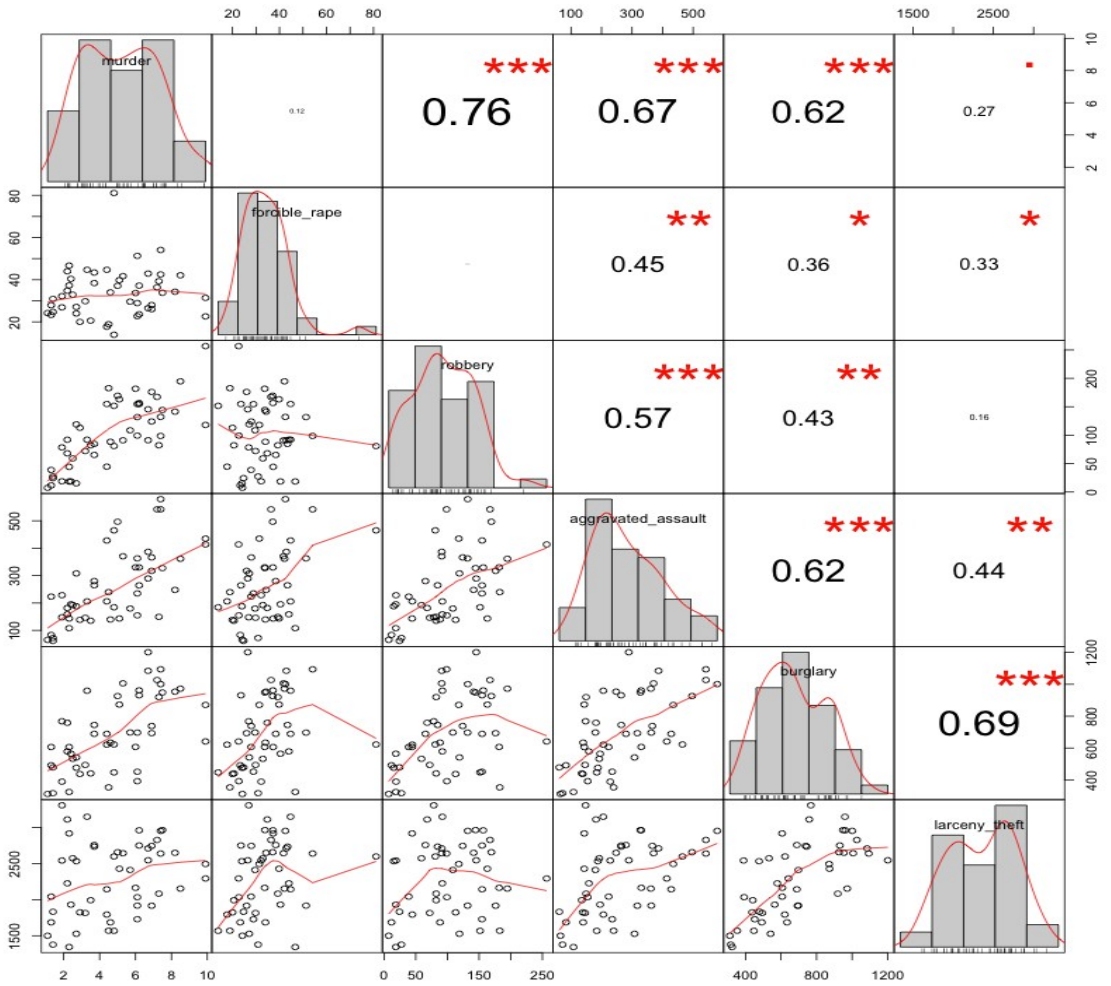
Adapted From Quick R: http://www.statmethods.net/advgraphs/layout.html

Example: US Crime Rates



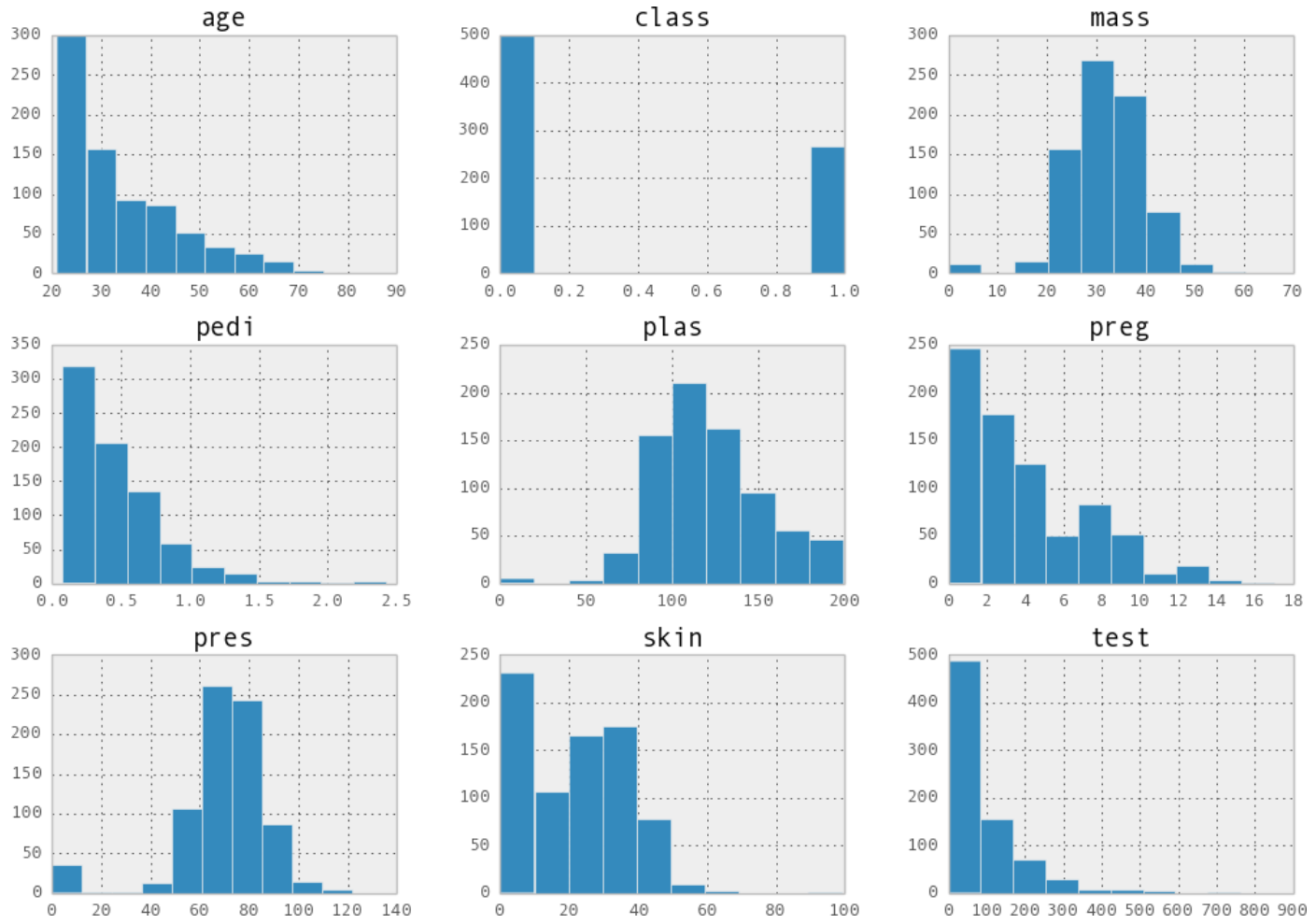
Adapted From R Bloggers: http://www.r-bloggers.com/quick-scatterplot-with-associated-histograms/

Example: US Crime Rates



Adapted From Getting Genetics Done: http://gettinggeneticsdone.blogspot.ie/2012/08/more-on-exploring-correlations-in-r.html

Histogram Matrix



Conclusion

- Visualisation workflow
  - Audience, Angle, Frame, Focus

- What relationships to look for?
  - Correlation ≠ Causation

- How many variables are we exploring?
- Are we looking to identify exact values?
- Are we looking to identify general patterns?

Useful Links

- https://www.guru99.com/tableau-charts-graphs-tutorial.html#3
- https://www.tutorialgateway.org/tableau/
- https://www.zapbi.com/blogs/chart_design_data_Visualisation_part_1/

Assignment 1 30%

Specification

You have been hired as a visualisation designer to design an effective dashboard providing insights into a dataset. As part of the visualisation process you will first explore the data and produce a dashboard useful for exploration, then you will set your editorial thinking and produce a dashboard with at least 3 insights from the data.

Marking scheme

1. Select a Dataset – 2%
2. Decide on an audience (user story) – 3%
3. Using Tableau Public, design a Dashboard that allows the exploration of the data - 8%
4. Using Tableau Public, design a Dashboard that shows at least three insights from the data - 12%
5. Show evidence of previous iterations or alternatives - 5%

Assignment 1 30%

Sample sources of data

- https://toolbox.google.com/datasetsearch
- https://archive.ics.uci.edu/ml/index.php
- https://data.gov.ie/
- https://public.tableau.com/en-us/s/resources
- Make Over Monday challenges

Setting up R

- Rstudio
- Anaconda =>
- Jupyter Notebooks with R kernel

Thanks To

- Marisa Llorens-Salvador, John McAuley, Colman McMahon and Brian Mac Namee for an earlier version of these lecture notes