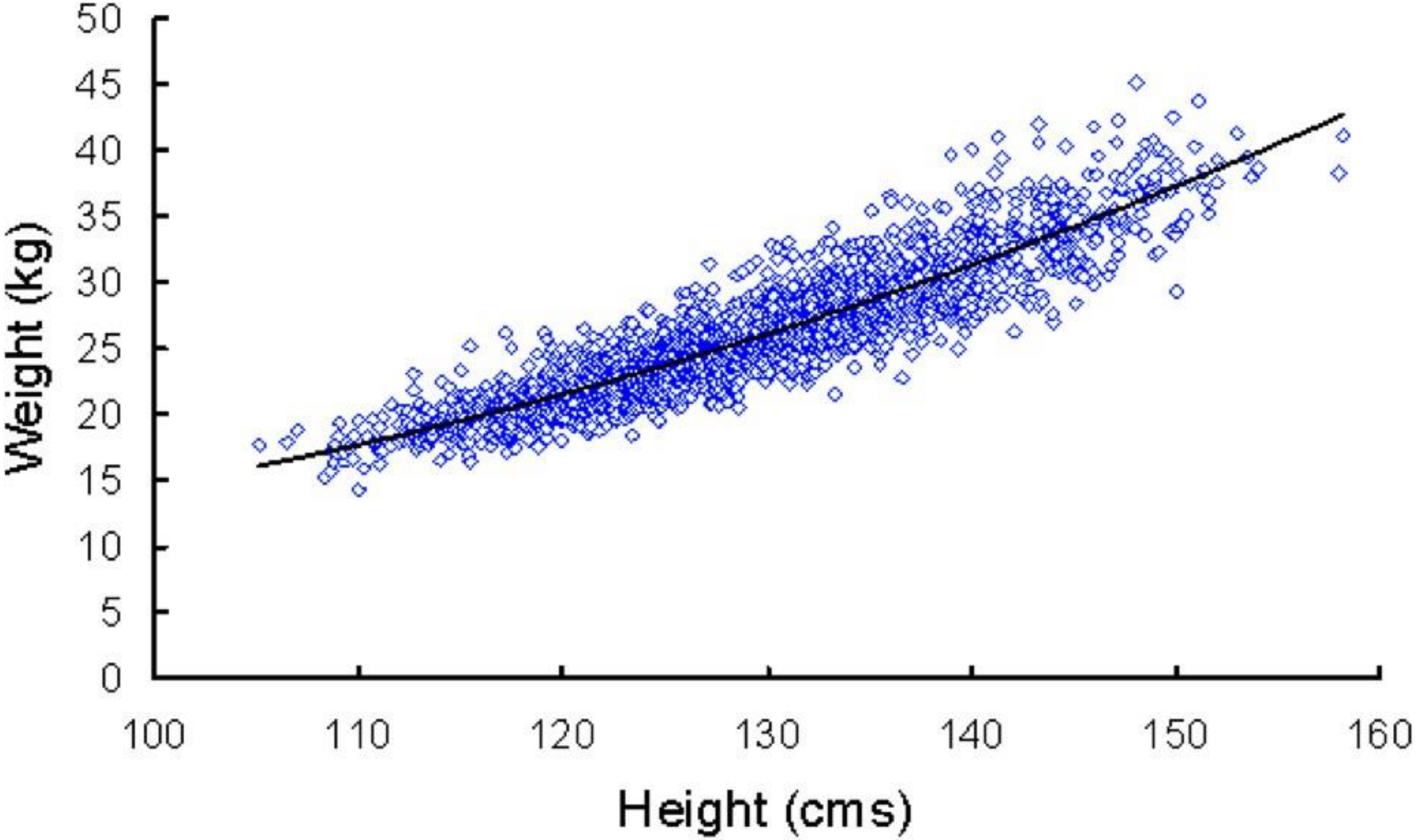
CORRELATION & CAUSATION

Statistics & Relationships

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

Height & Weight



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

Correlation & Causation

Correlation

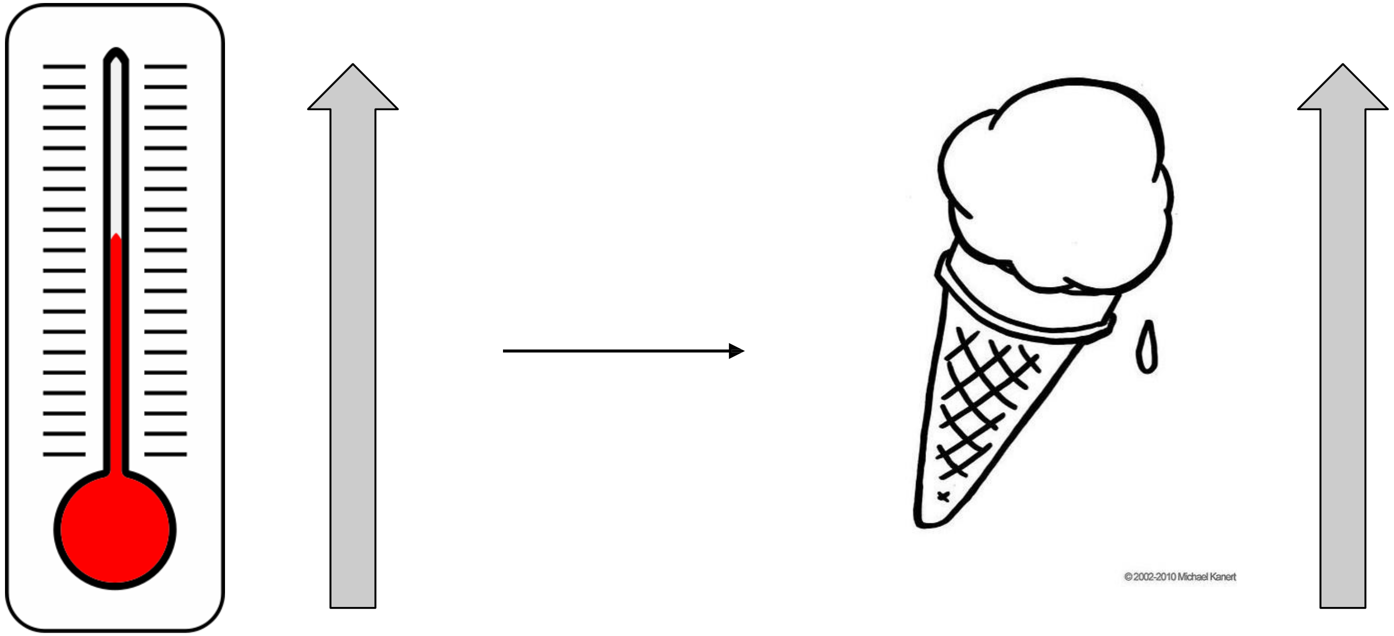
- "A statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables."

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



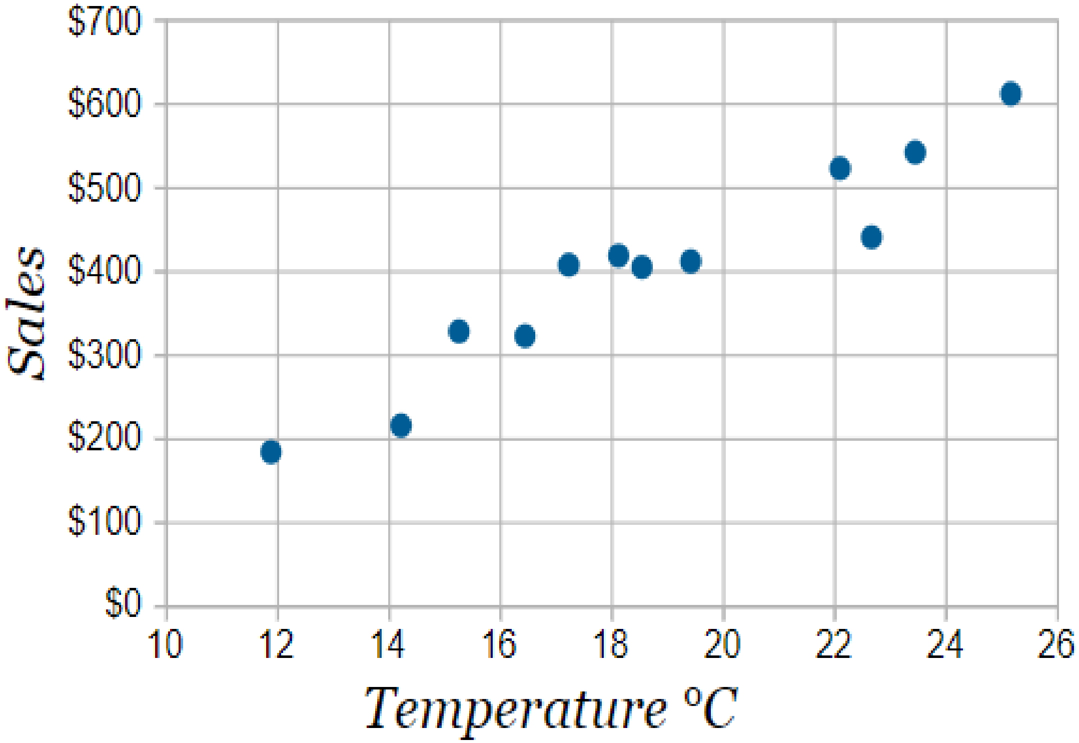
Independent variable

Dependent variable

Correlation & Causation

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

Icecream



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

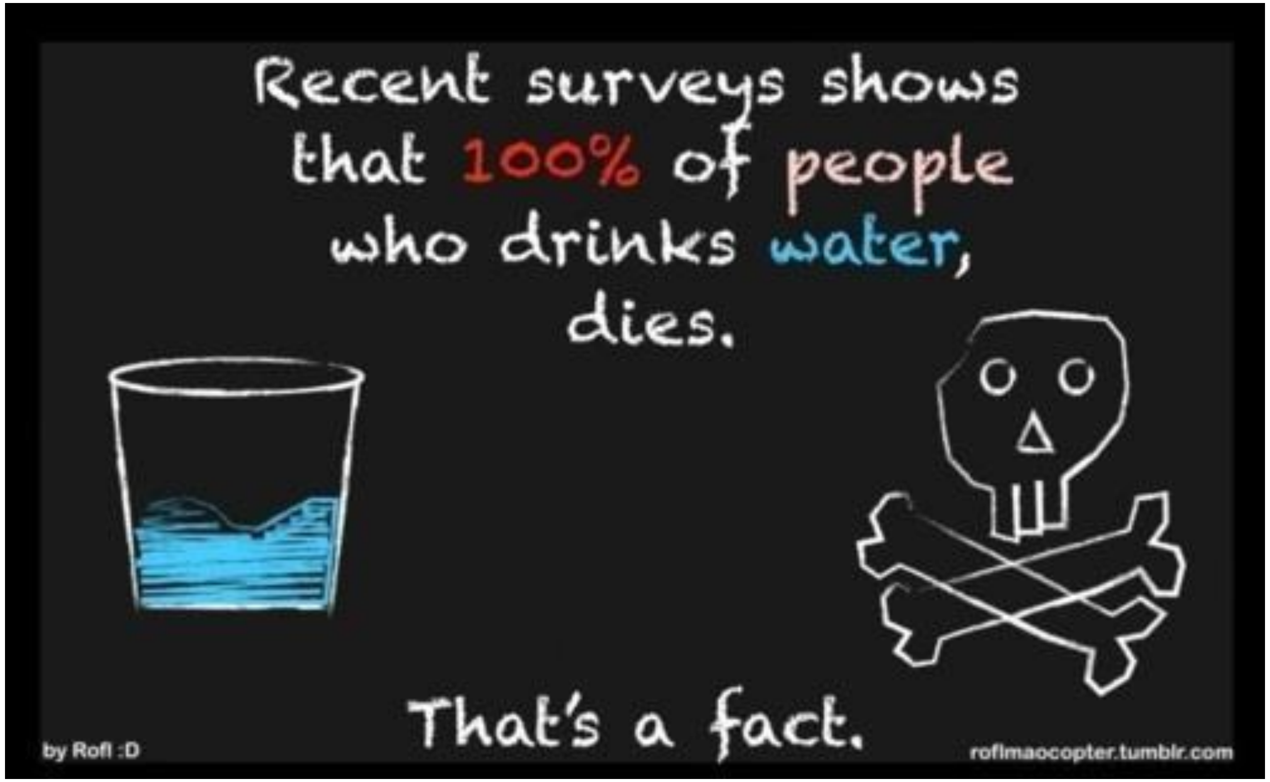
What Relationships To Look For?

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)

Correlation & Causation

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


Recent surveys shows that 100% of people who drinks water, dies.
That's a fact.
by Rofl :D
roflmaocopter.tumblr.com

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

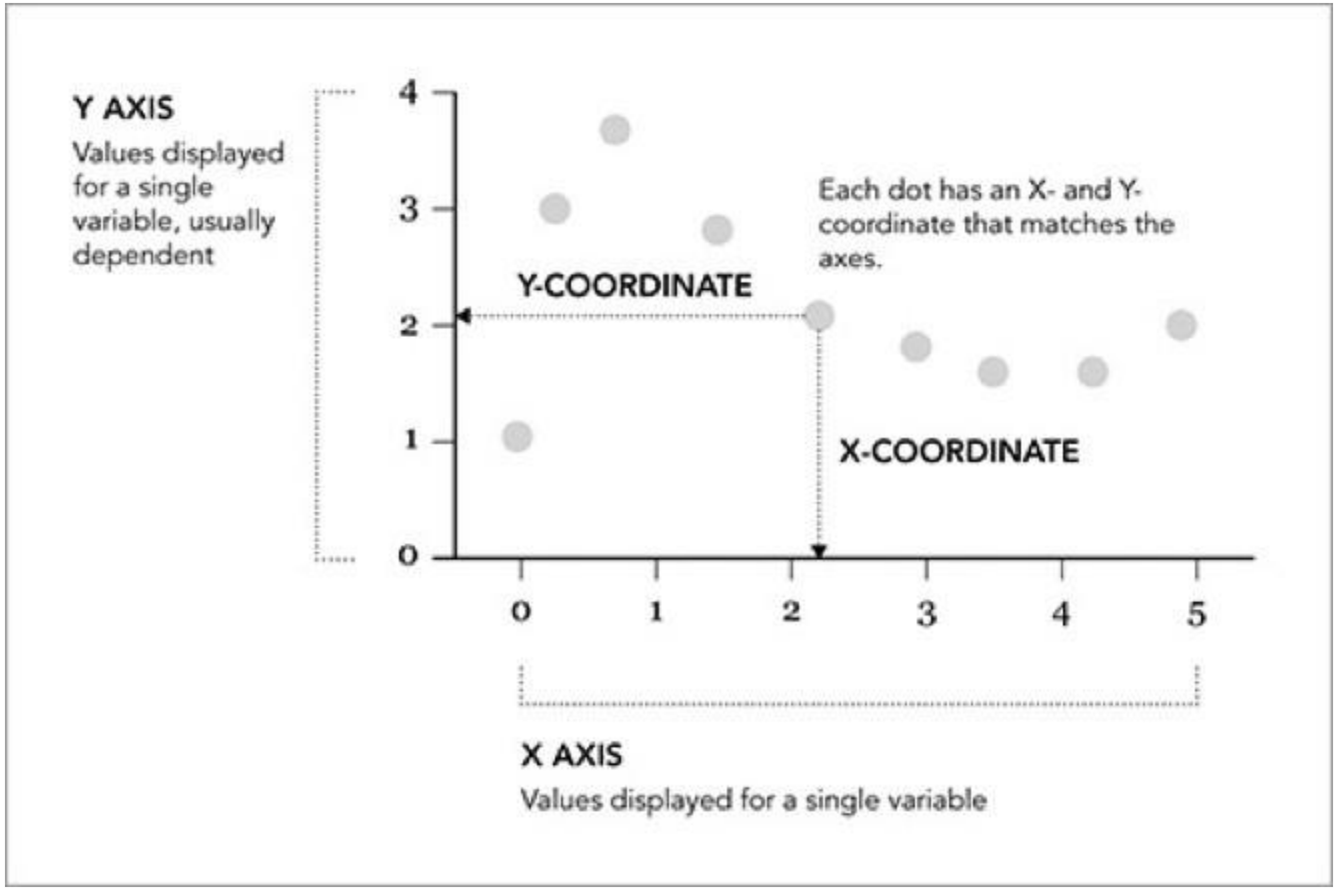
Finding Correlation

- 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

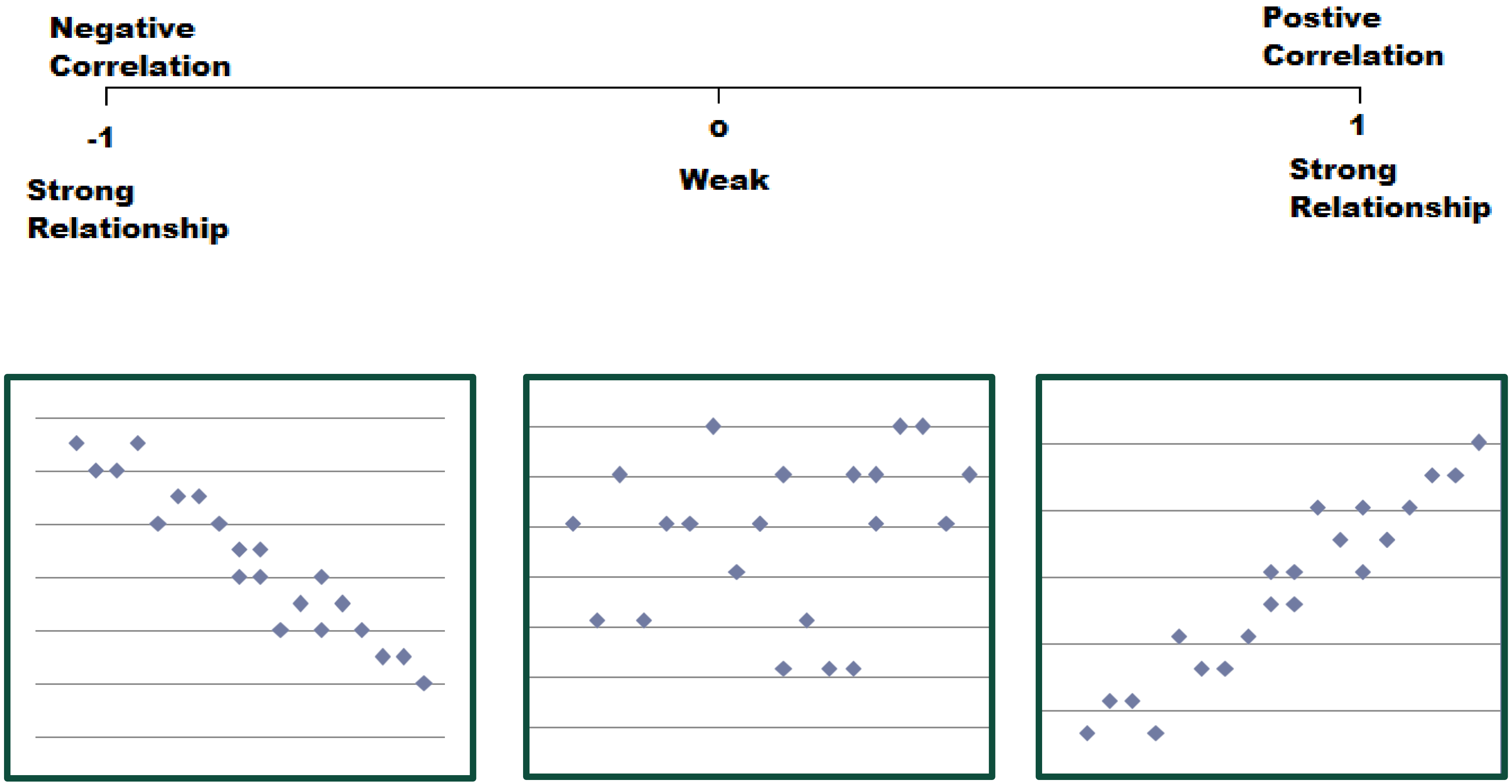
Simple Scatter Plot



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



Negative Correlation

Postive Correlation

-1

0

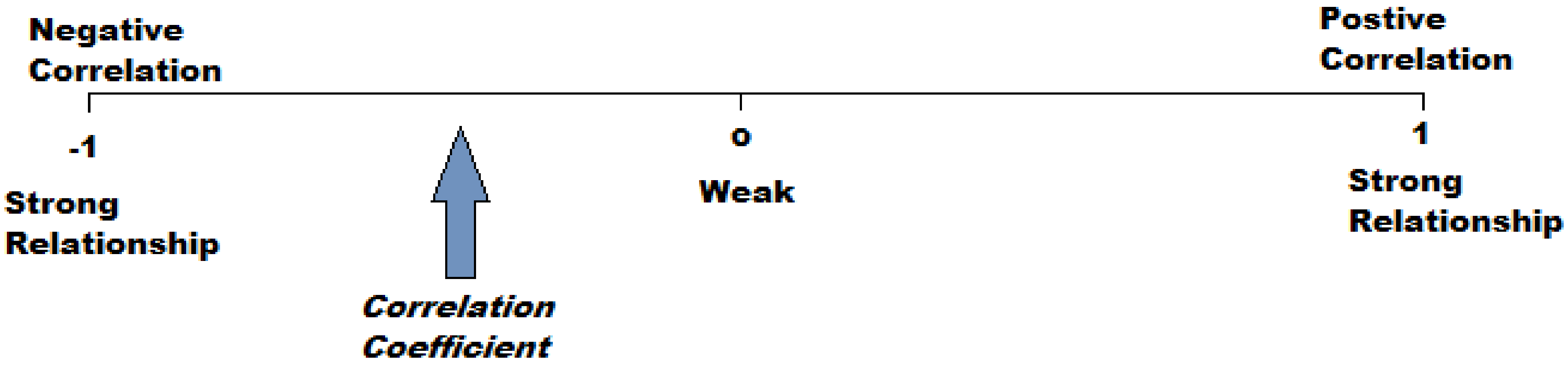
1

Strong Relationship

Weak

Strong Relationship

Simple Scatter Plot



Negative Correlation

Postive Correlation

-1

Strong Relationship

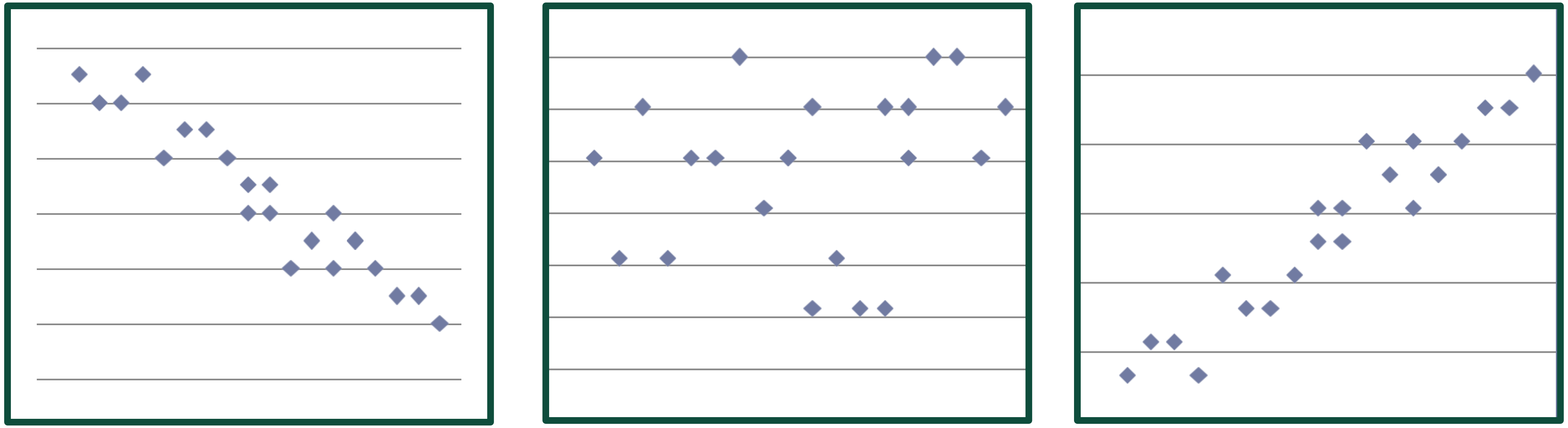
0

Weak

1

Strong Relationship

Correlation Coefficient

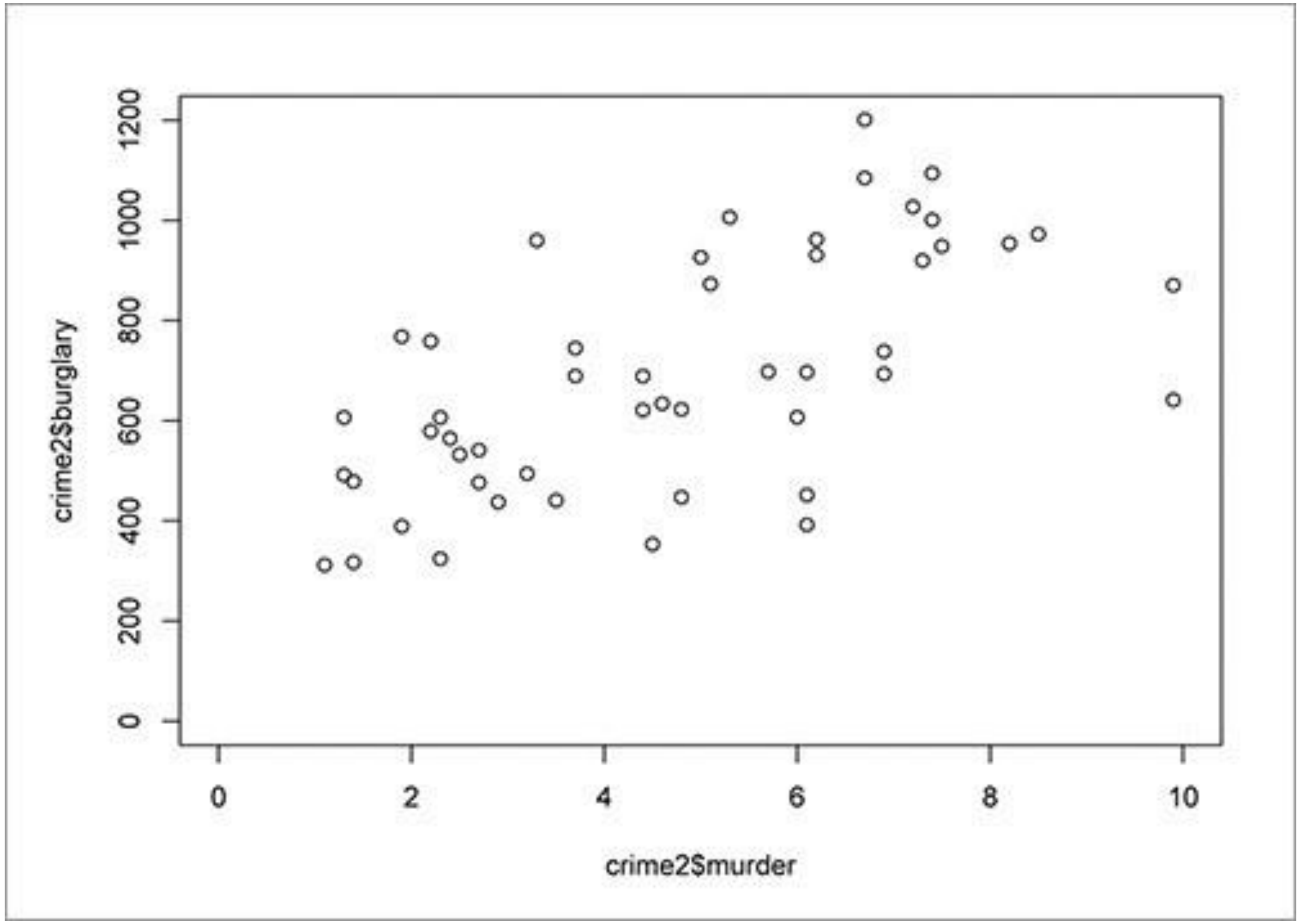


Simple Scatter Plot

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

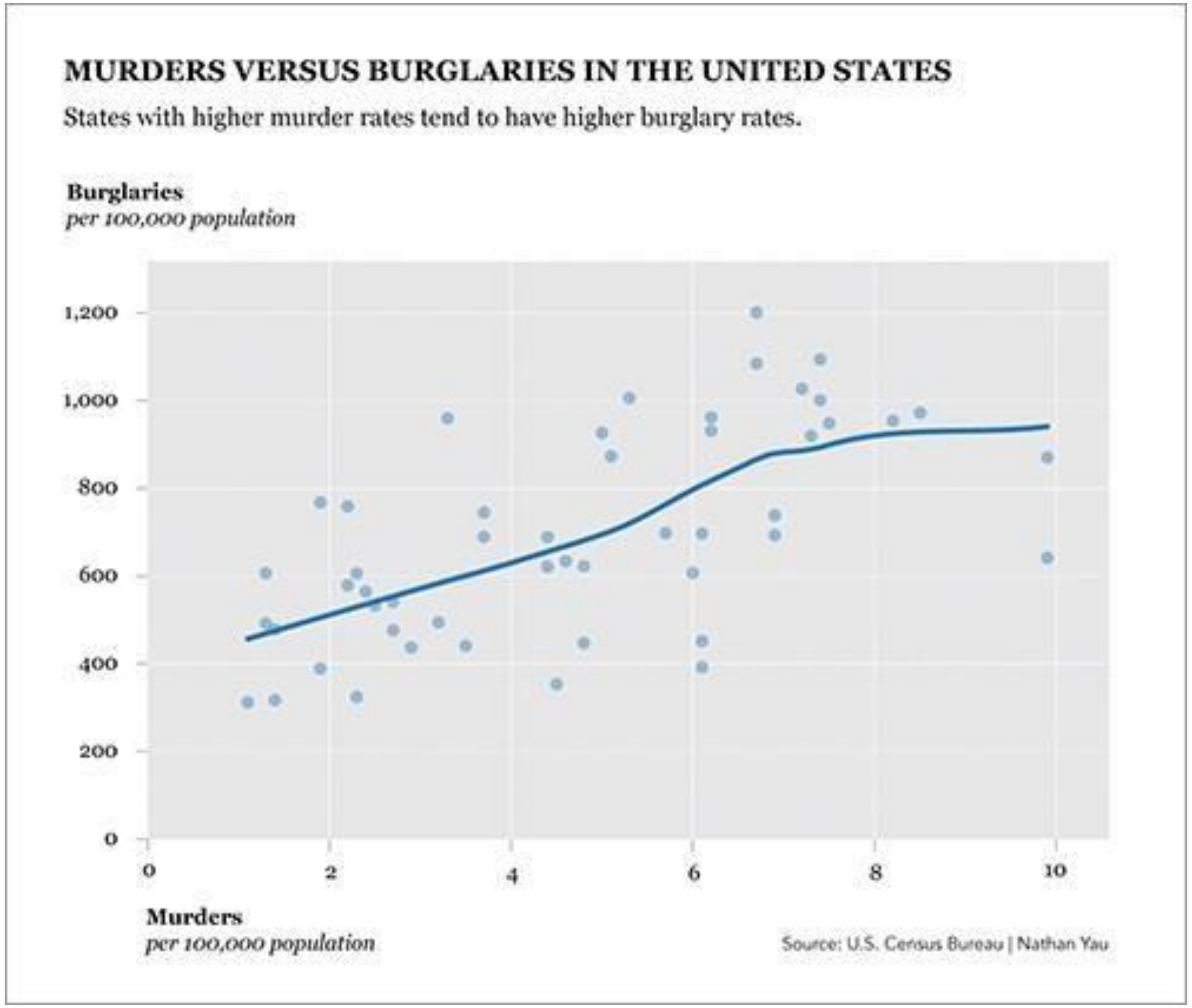
- Composition
  - X- independent variable
  - Y- dependent variable
  - 1:1 aspect ratio
  - No need to start at 0

Example: US Crime Rates



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

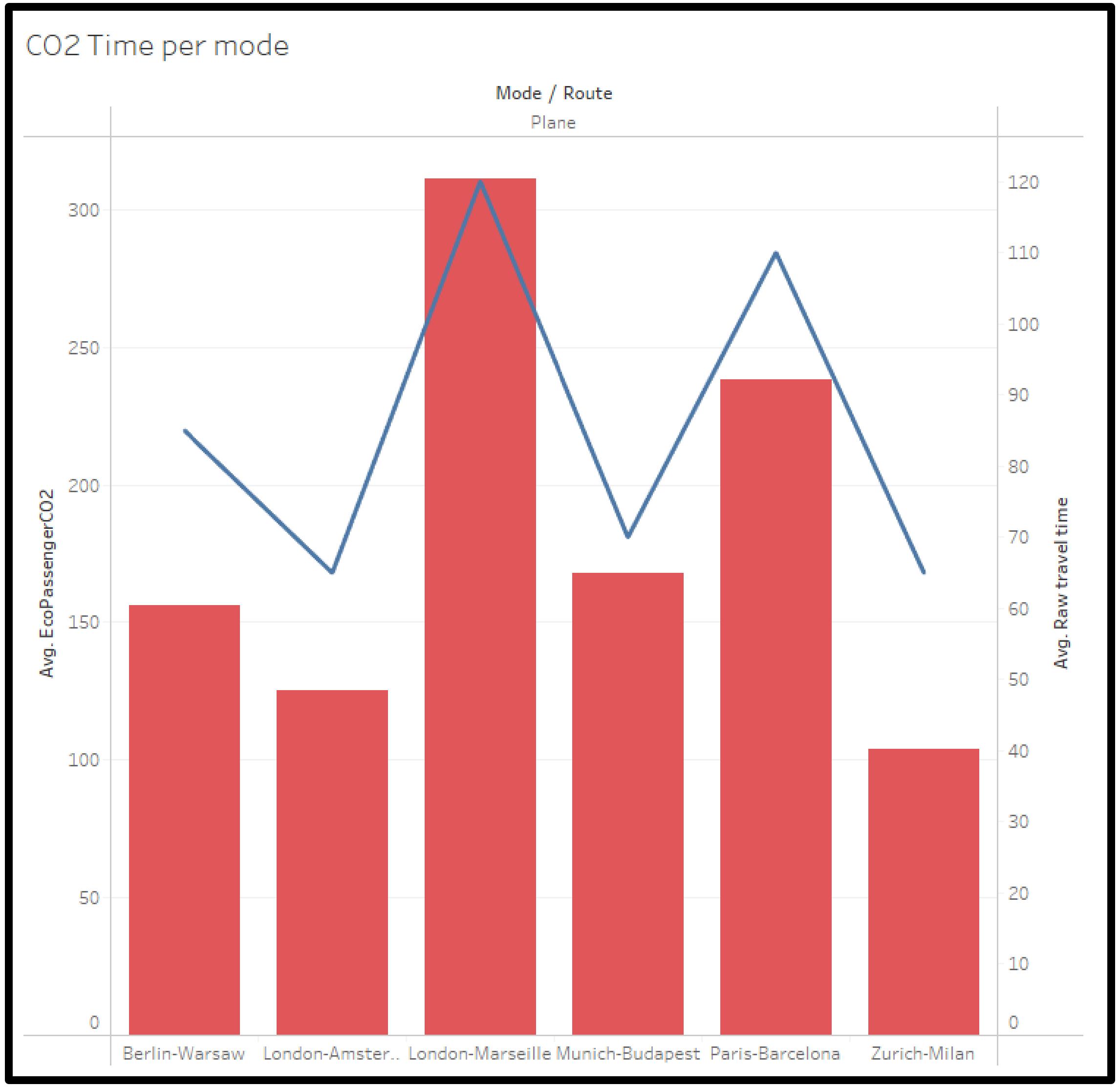
Example: US Crime Rates



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

LINE COLUMN CHARTS

Line Column Charts



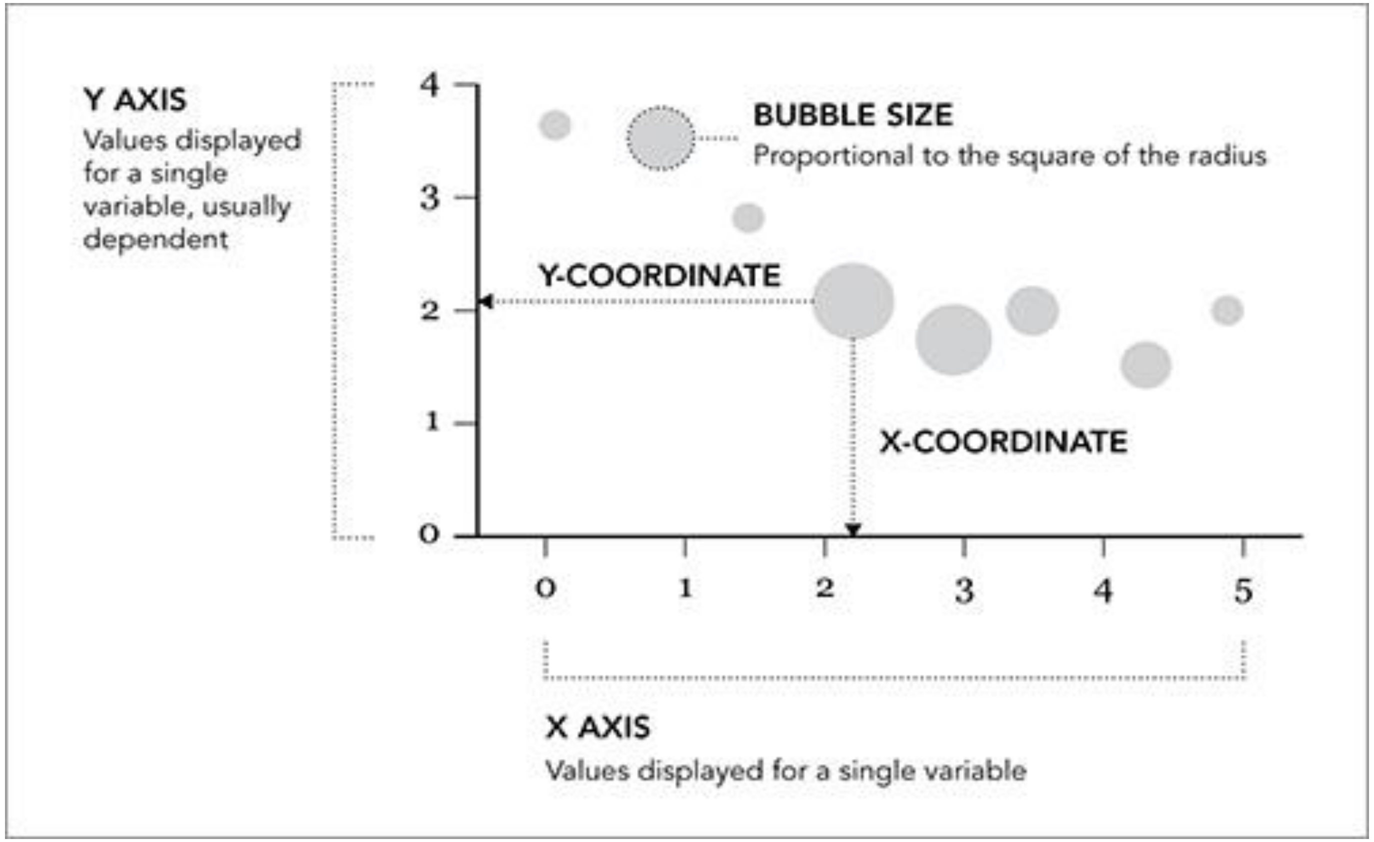
- Easily illustrates the relationships between two variables with different magnitudes and scales of measurement

- Note secondary axis

BUBBLE PLOTS

Bubble plots

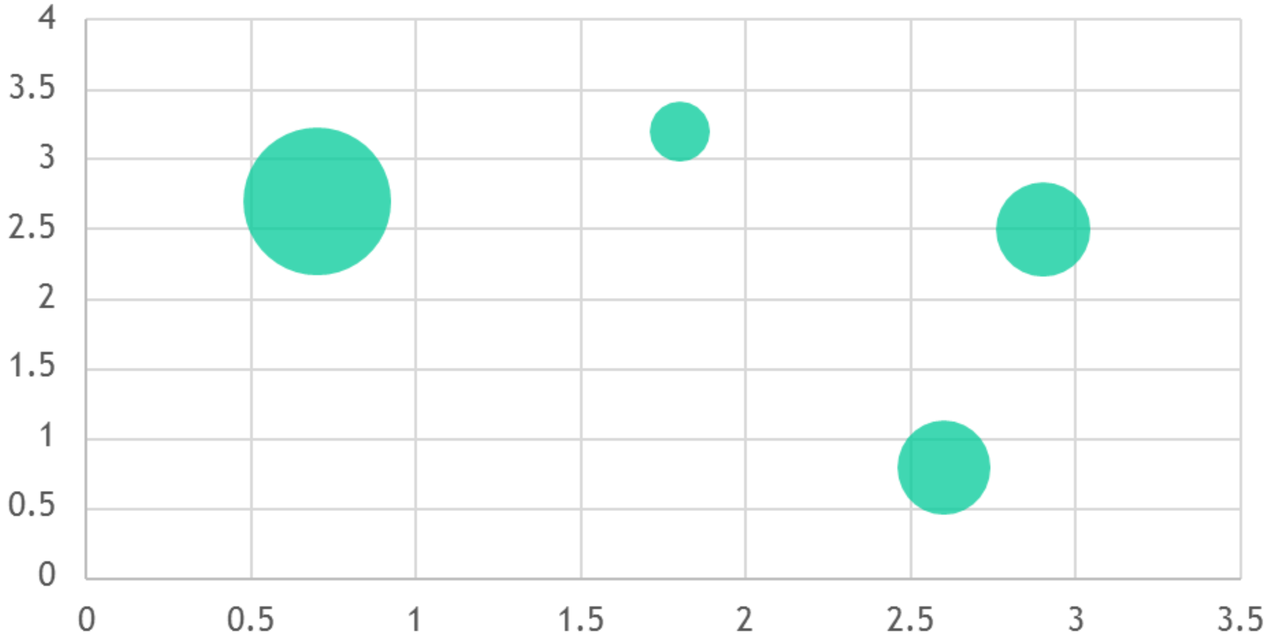
- A bubble plot can be defined as a 3D scatterplot
  - The value of an additional variable is represented through the size of the dots.



Y AXIS
Values displayed for a single variable, usually dependent
BUBBLE SIZE
Proportional to the square of the radius
Y-COORDINATE
X-COORDINATE
0
1
2
3
4
0
1
2
3
4
5
X AXIS
Values displayed for a single variable

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.



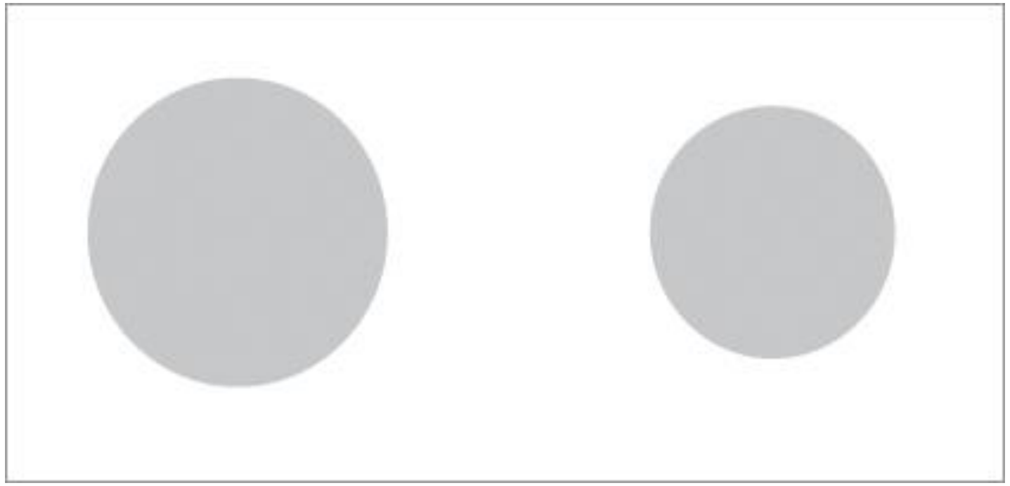
Bubble plots - Composition

- 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

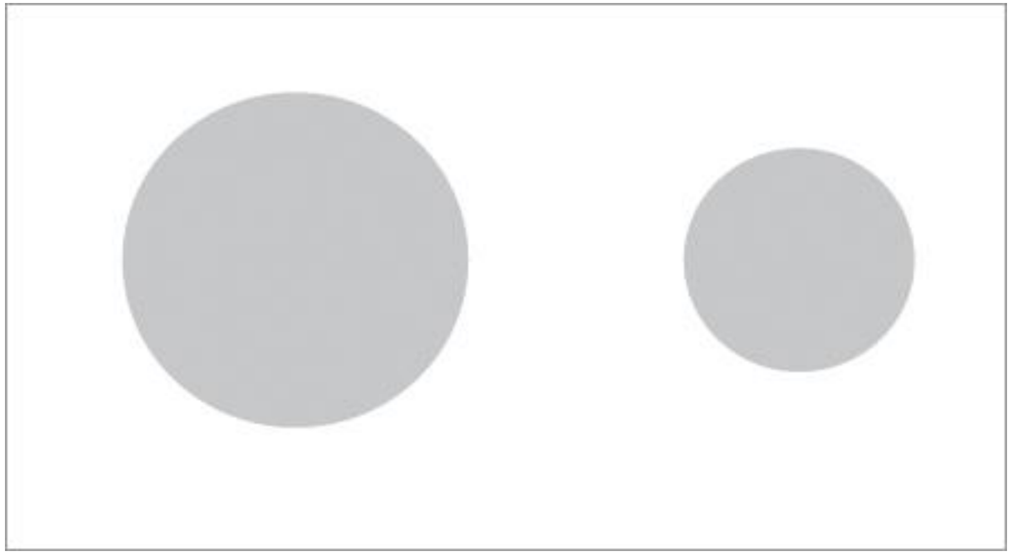
Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics. Nathan Yau. 2011

Bubble Plots

Sized by Area

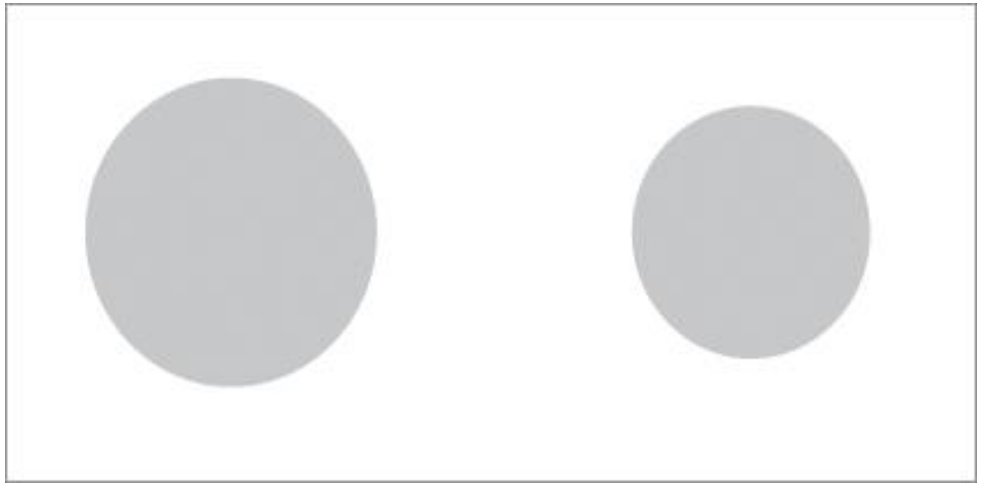


Sized by Radius

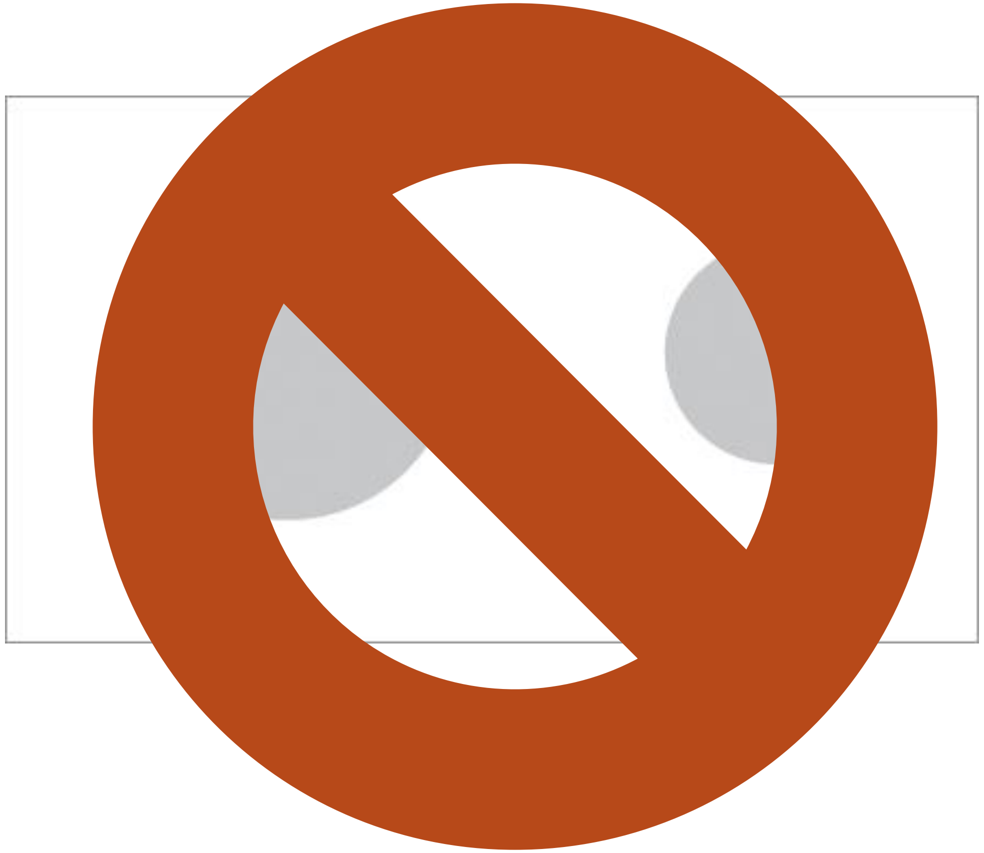


Bubble Plots

Sized by Area

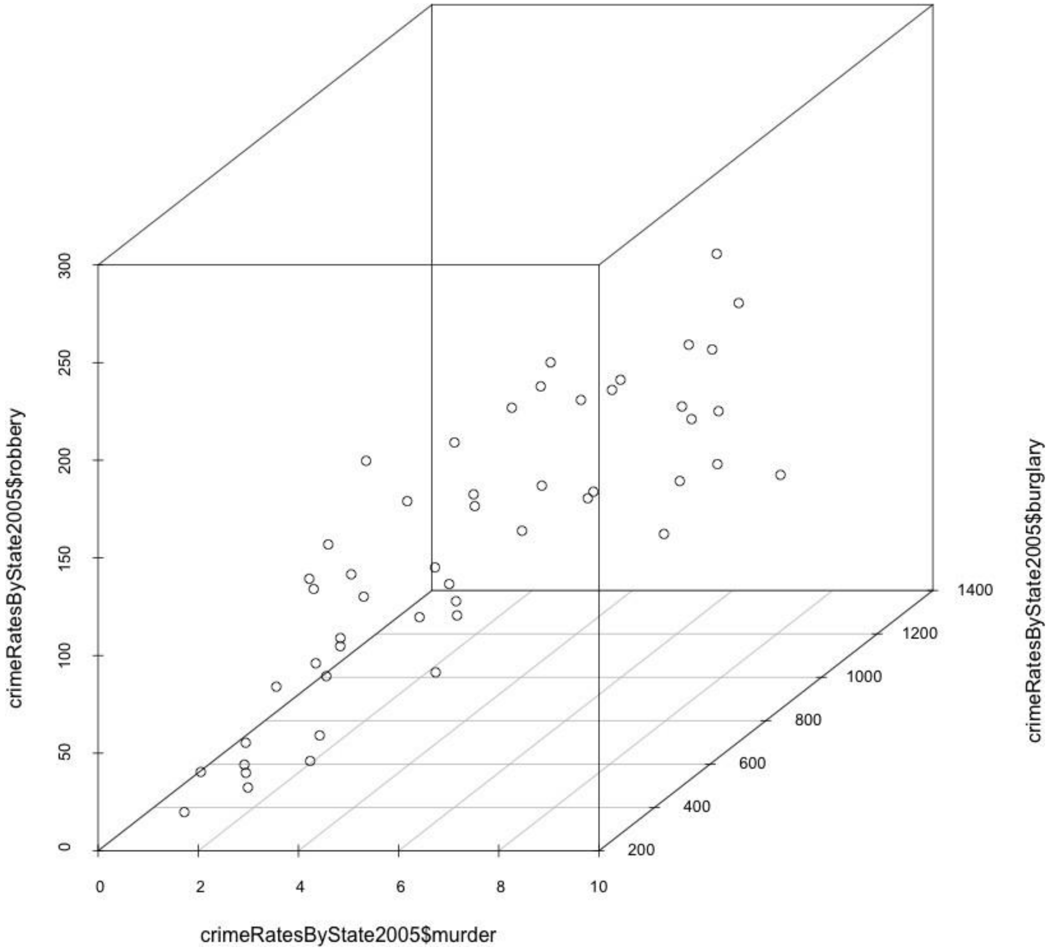


Sized by Radius



3D SCATTER PLOTS

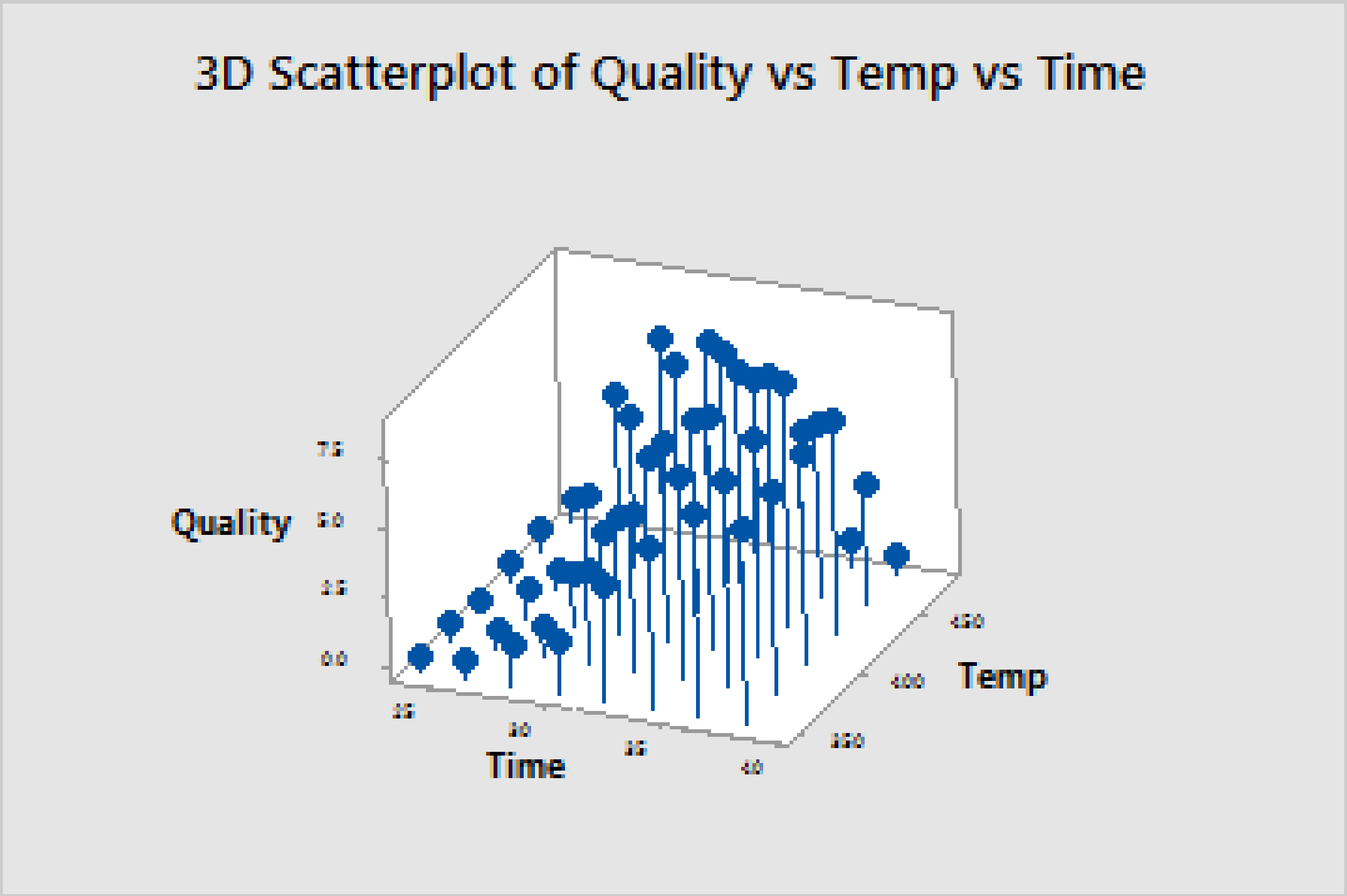
3D scatter Plots



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

3D scatter Plots

- optimal time and temperature for heating a frozen dinner.



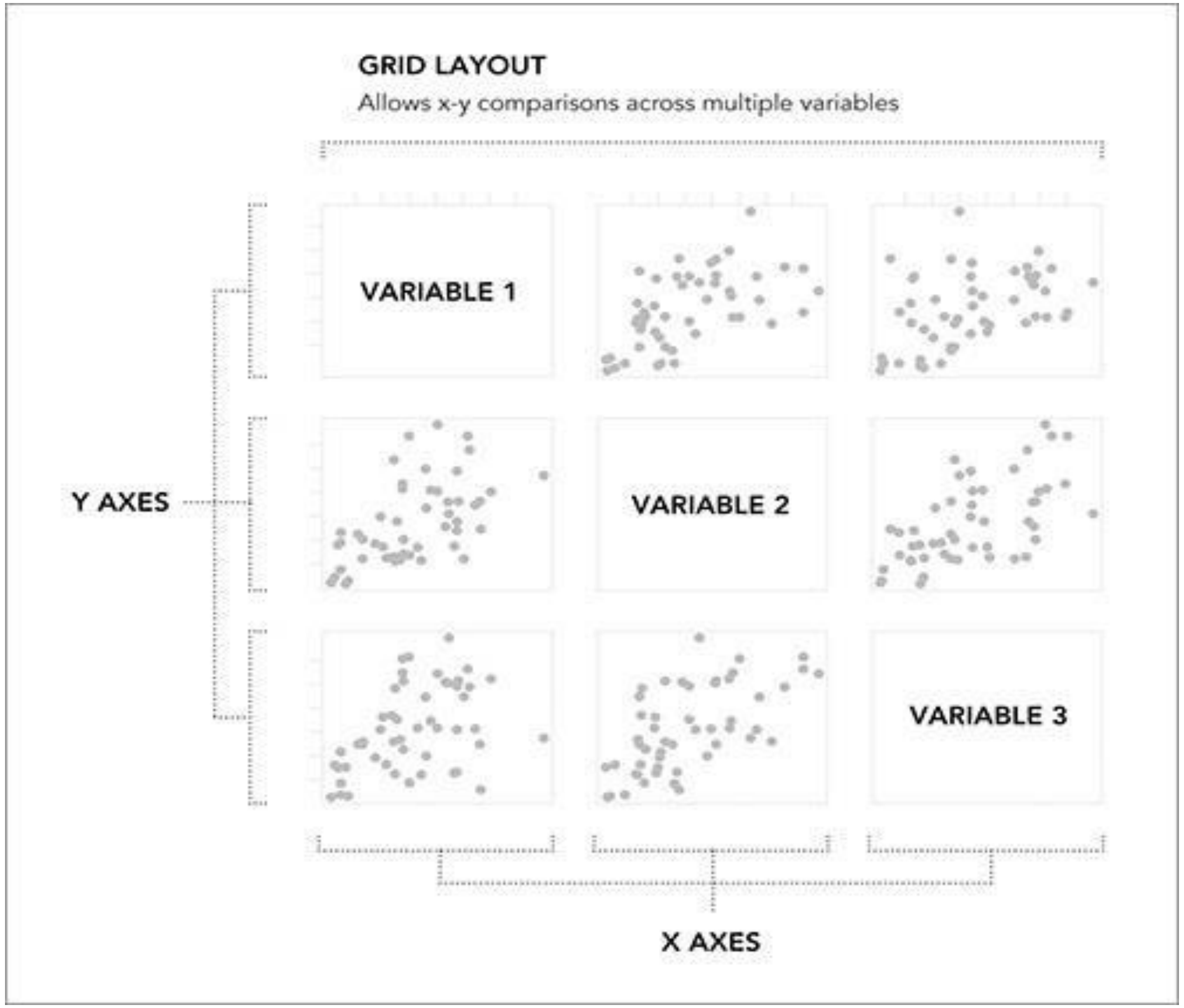
https://support.minitab.com/en-us/minitab/19/help-and-how-to/graphs/3d-scatterplot/before-you-start/example/

EXTRA DIMENSIONS

Exploring Even More Variables

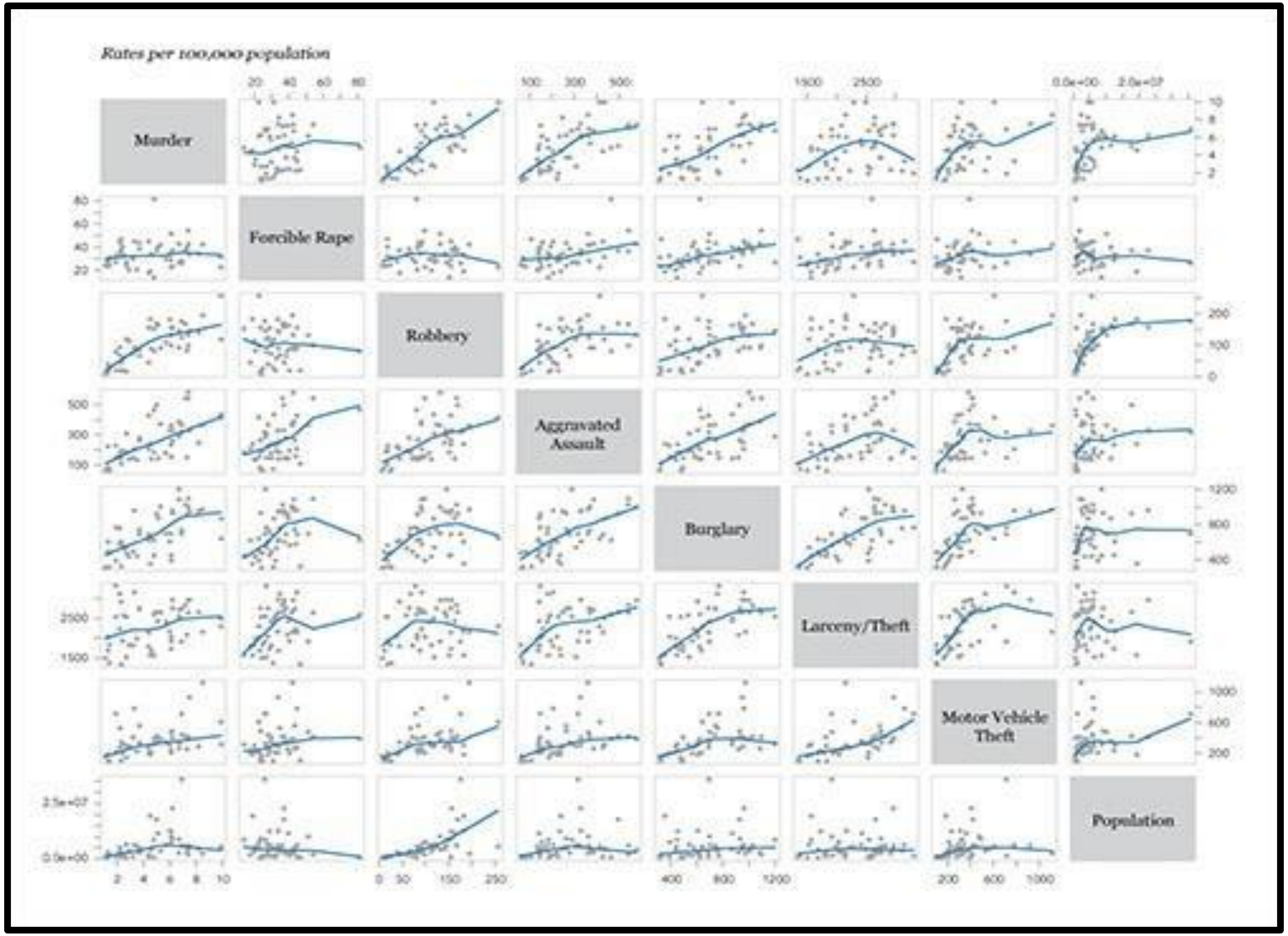
- You can plot every possible pair with a scatter plot matrix to compare all variables
- It’s usually a square grid with all variables on both the vertical and horizontal
- Each column represents a variable on the horizontal axis, and each row represents a variable on the vertical axis
- This provides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

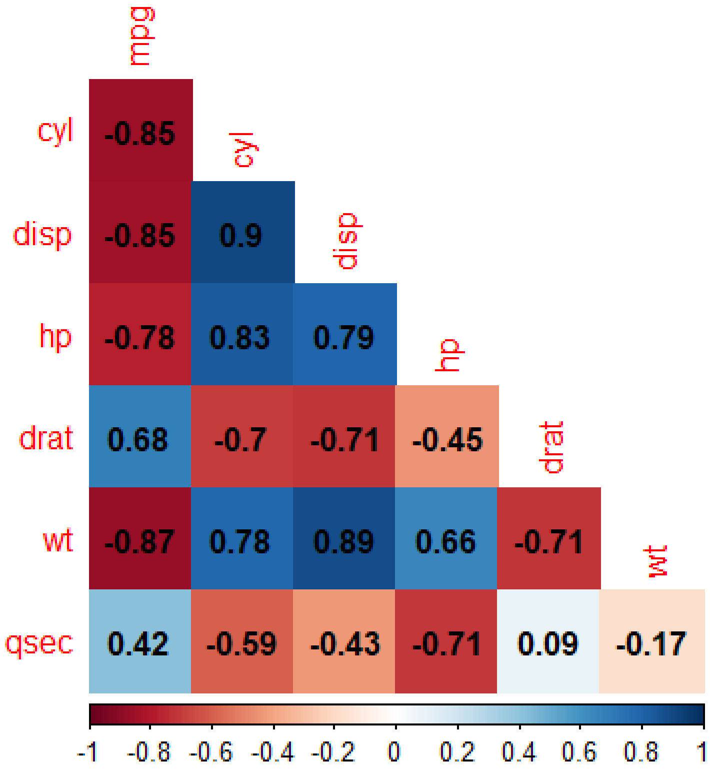
- A heat map displays quantitative values at the intersection between two categorical dimensions

- Two categorical axis with all possible values

- Each cell is colour coded to represent a quantitative value for each combination of category pairing

Heatmap Coloured Correlation Matrix

- How do the values in the columns in mtcars correlate to each other?
- The Pearson Correlation assigns a value of between -1 and 1 to indicate that the values are negatively correlated, not correlated (0) or positively correlated (heading for 1).
- https://lost-stats.github.io/Presentation/Figures/heatmap_colored_correlation_matrix.html#heatmap-colored-correlation-matrix



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

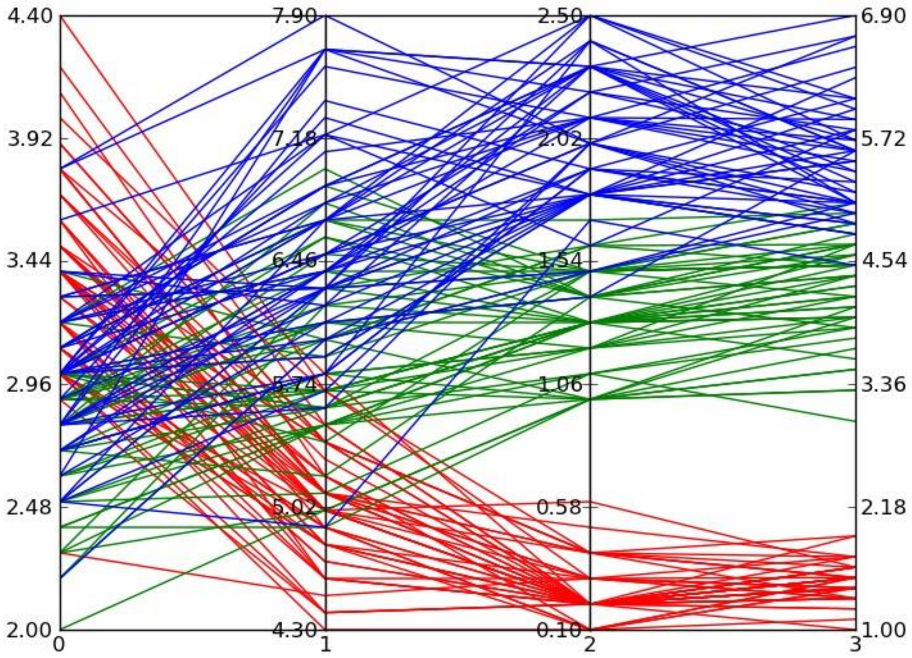
- Composition:
  - Logical sorting and sub-grouping can aid readability
  - Colour scale

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

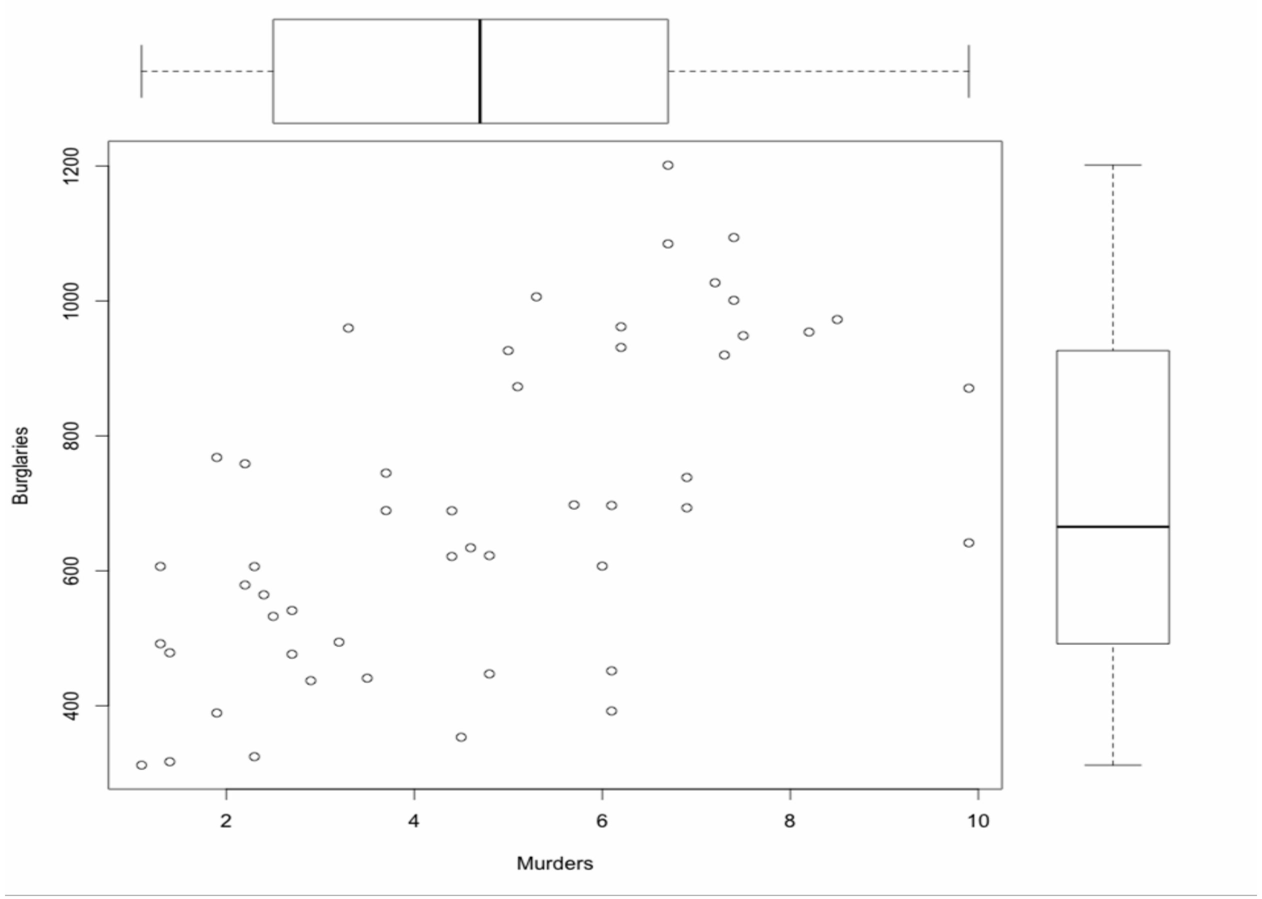
EXTENSIONS TO SCATTER PLOTS

Playing With Scatter Plots

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

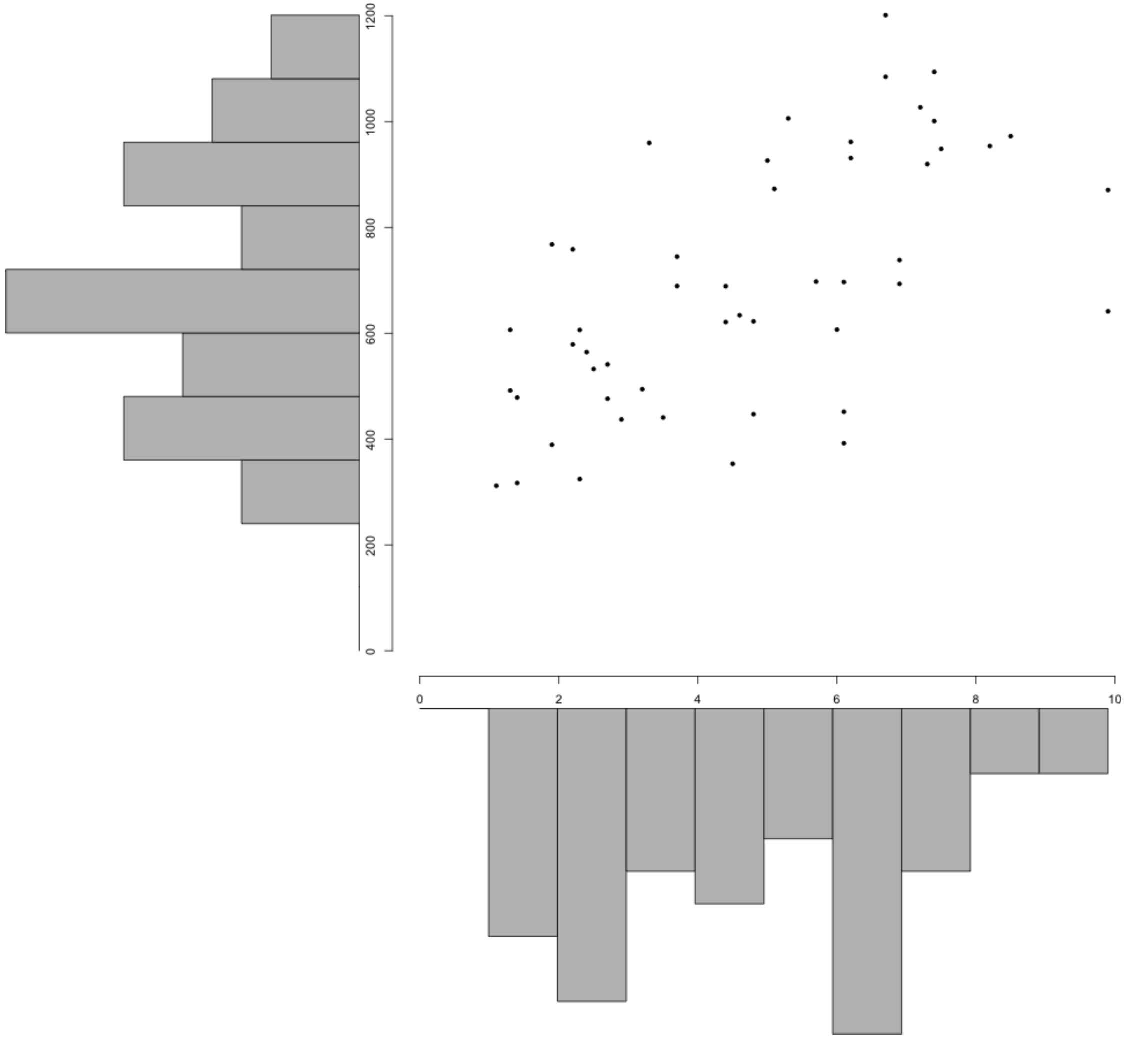
- We can augment scatter plots with other interesting things to show more information

Example: US Crime Rates



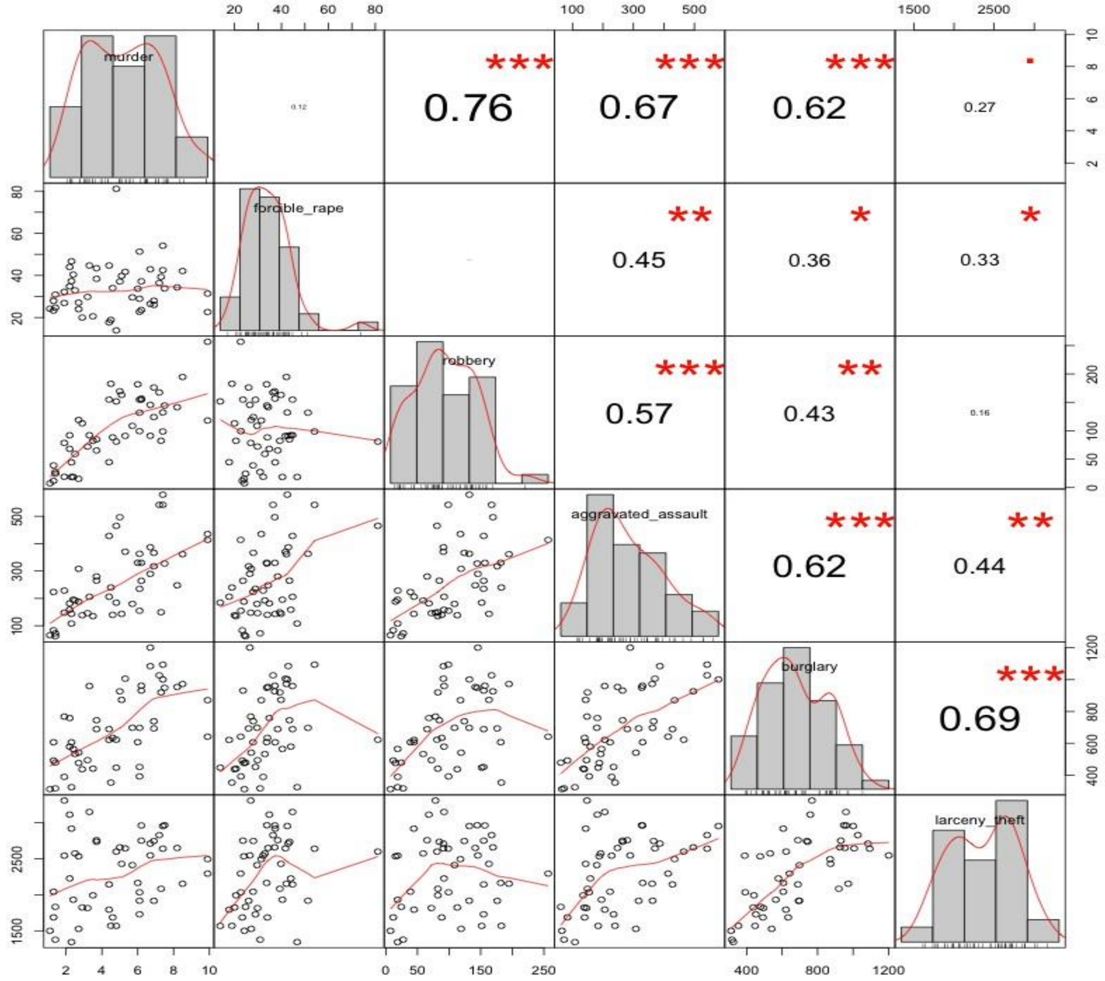
Adapted From Quick R: http://www.statmethods.net/advgraphs/layout.html

Example: US Crime Rates



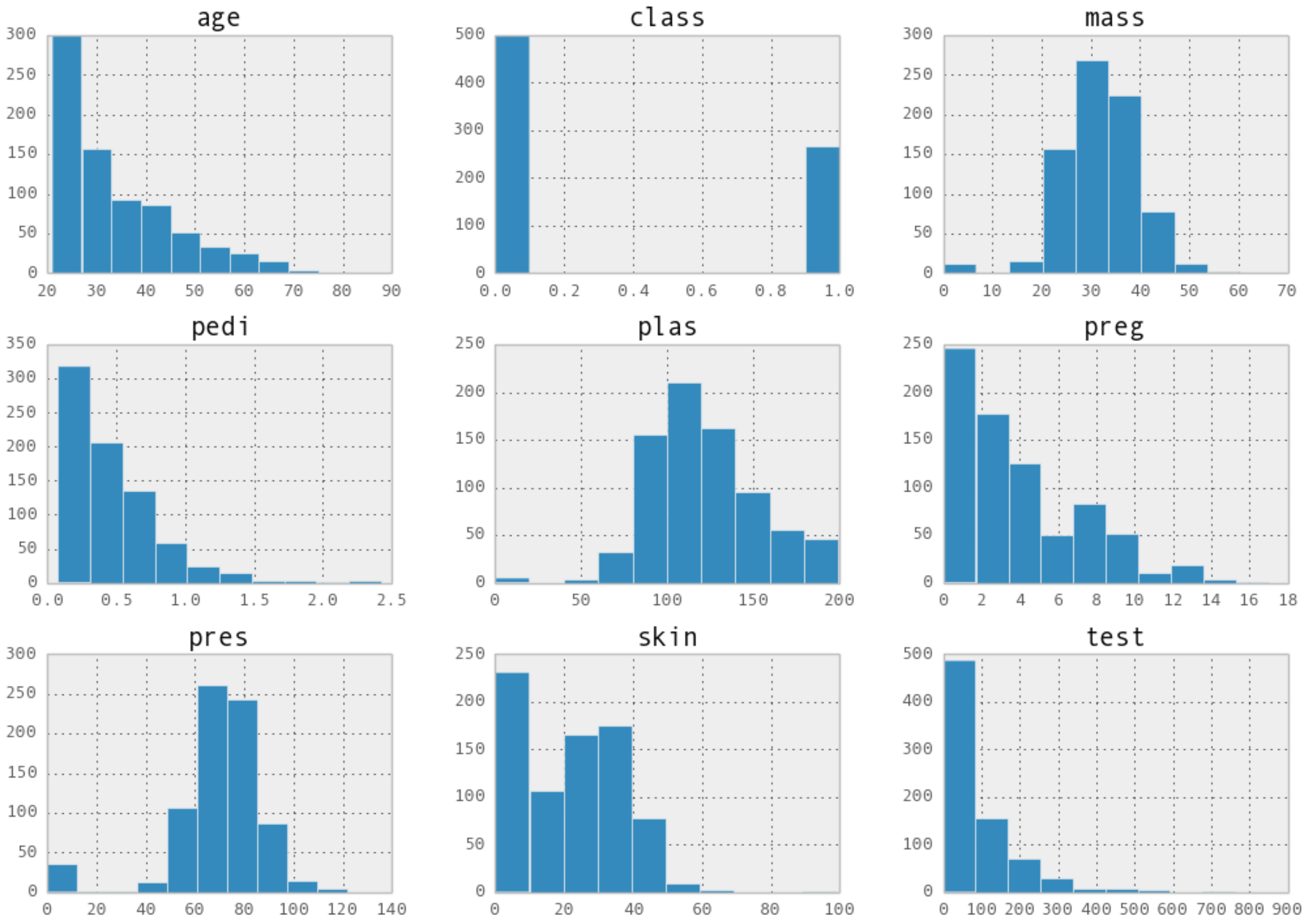
Adapted From R Bloggers: http://www.r-bloggers.com/quick-scatterplot-with-associated-histograms/

Example: US Crime Rates



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

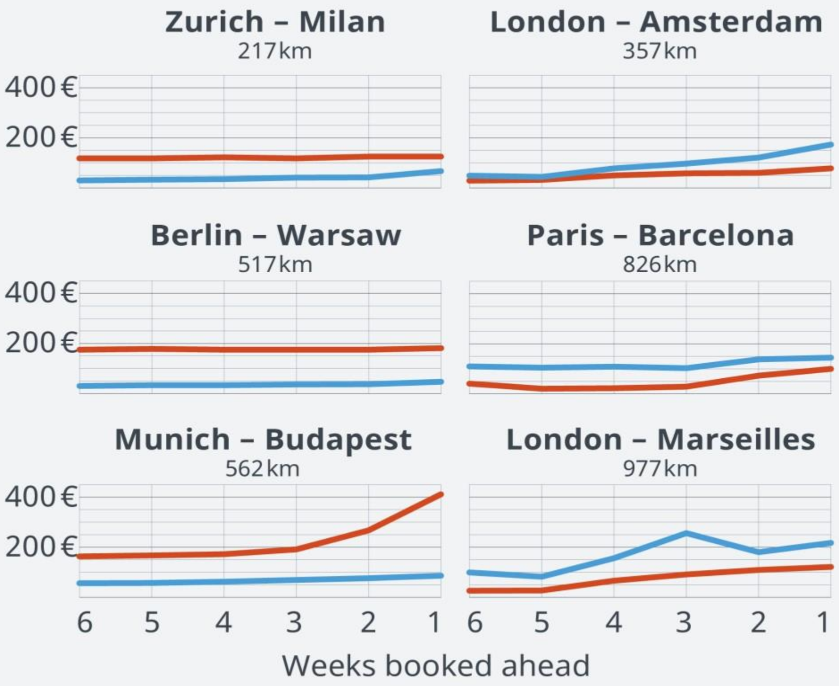
Histogram Matrix



Make Over Monday Exercise

Planes not always cheaper than trains

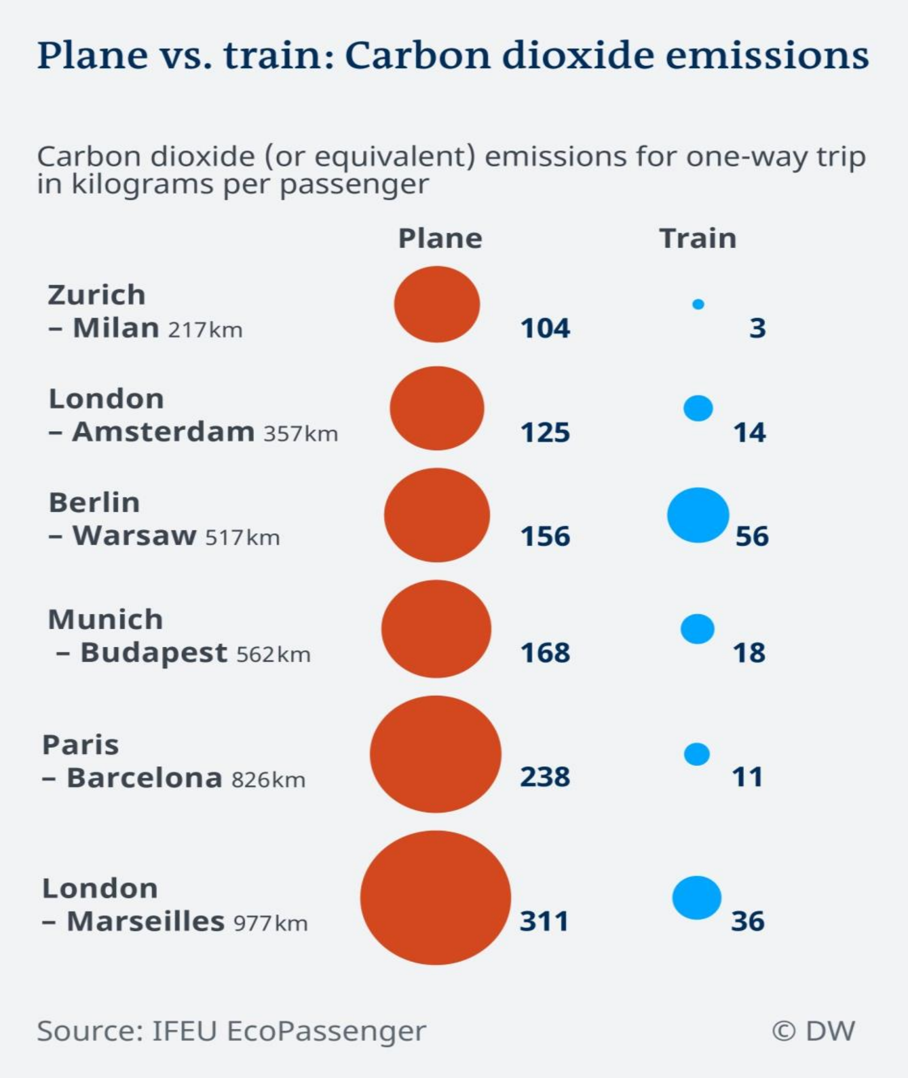
Ticket prices for planes and trains on different routes



Weekly average price in euros for cheapest daily direct one-way connection

Source: Google Flights, Trainline, DW analysis © DW

Make Over Monday Exercise



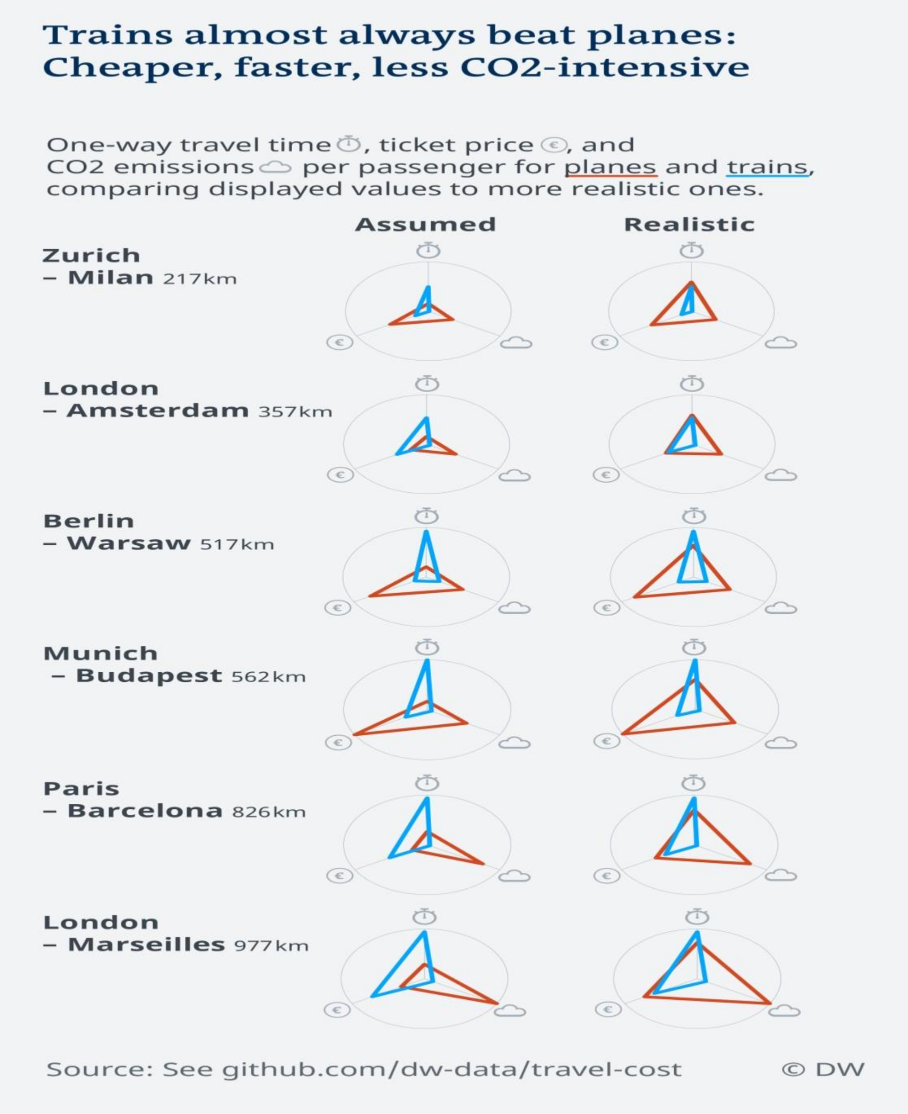
Plane vs. train: Carbon dioxide emissions

Carbon dioxide (or equivalent) emissions for one-way trip in kilograms per passenger

Source: IFEU EcoPassenger

© DW

Make Over Monday Exercise



Trains almost always beat planes: Cheaper, faster, less CO2-intensive

One-way travel time ⏱, ticket price €, and CO2 emissions ☁ per passenger for planes and trains, comparing displayed values to more realistic ones.

Source: See github.com/dw-data/travel-cost © DW

Thanks To

- Cathy Ennis, Marisa Llorens-Salvador, John McAuley, Colman McMahon and Brian Mac Namee for earlier versions of these lecture notes

Evaluating Charts

Evaluating charts

- What is the chart type?
- Is it appropriate to the data? The message?
- Did you understand it immediately? In a few minutes?
- How many dimensions are there?
- How easily can you make them out?
- Are they well labelled?
- Are they suitable to the audience?

ACCENT Principles for effective graphical display

- The essence of a graph is the clear communication of quantitative information.
- The ACCENT principles emphasize, or accent, six aspects that determine the effectiveness of a visual display for portraying data.

1. Apprehension:
    - The Ability to correctly perceive relations among variables.
    - Does the graph maximize apprehension (understanding) of the relations among variables?

ACCENT Principles for effective graphical display

2. Clarity:
   - Ability to visually distinguish all the elements of a graph.
   - Are the most important elements or relations visually most prominent?
3. Consistency:
   - Ability to interpret a graph based on similarity to previous graphs.
   - Are the elements, symbol shapes and colors consistent with their use in previous graphs?

ACCENT Principles for effective graphical display

4. Efficiency:
    - Ability to portray a possibly complex relation in as simple a way as possible.
    - Are the elements of the graph economically used?
    - Is the graph easy to interpret?

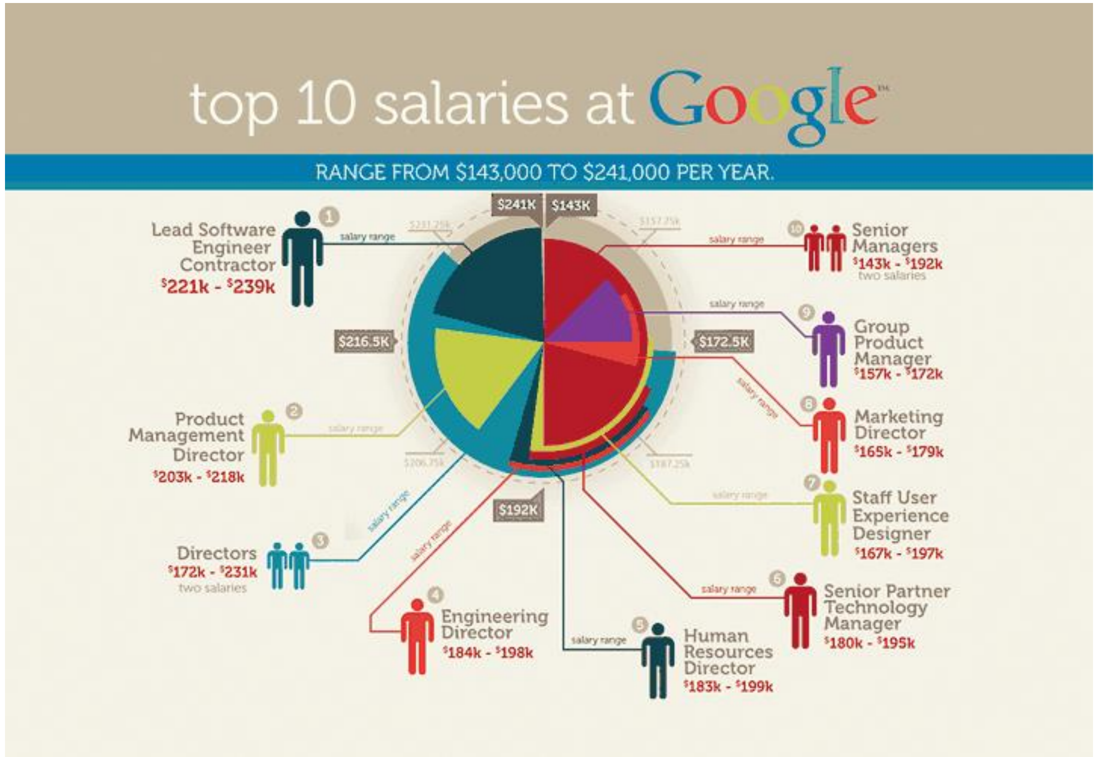
ACCENT Principles for effective graphical display

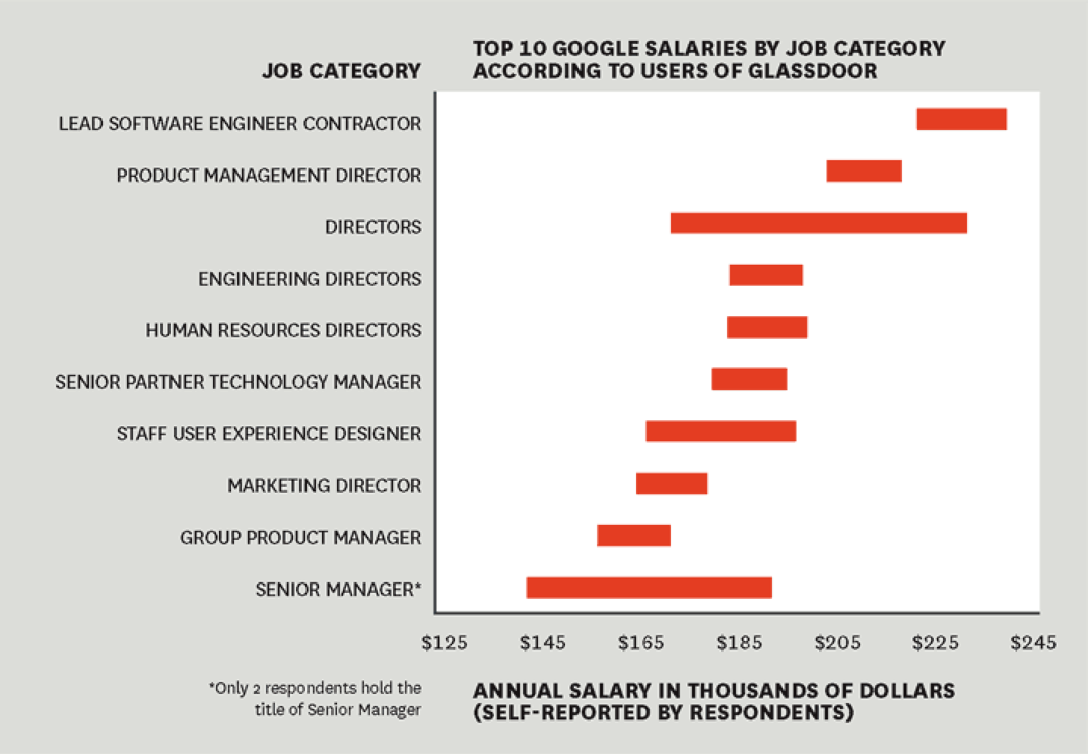
5. Necessity:
   - The need for the graph, and the graphical elements.
   - Is the graph a more useful way to represent the data than alternatives (table, text)?
   - Are all the graph elements necessary to convey the relations?
6. Truthfulness:
   - Ability to determine the true value represented by any graphical element by its magnitude relative to the implicit or explicit scale.
   - Are the graph elements accurately positioned and scaled?

Accent reminder

- A – Apprehension / understanding
- C - Clarity
- C - Consistency
- E - Efficiency
- N - Necessity
- T - Truthfulness

Same information – evaluate?





Source: http://www.jobvine.co.za/ Vevox Session: 118-967-269

How to spot a misleading graph

https://youtu.be/E91bGT9BjYk

Moving towards your project

Establish a Big idea

So what's the story?

Formulating your Big Idea

Getting from data to visualisation

Big Idea

- The Big Idea has three components:

1. It must articulate your unique point of view.
2. It must convey what's at stake.
3. It must be a complete sentence.

- NOTE: You may need to explore your data before you decide on your Big Idea!!!

The Big Idea Formulation

- Identify a project you are working on, where you need to communicate in a data-driven way. Reflect upon, and fill out the following:
- **Who is your audience?:**
  - List the primary groups or individuals to whom you will be communicating.
  - If you had to narrow that down to a single person, who would it be?
  - What does your audience care about?
  - What action does your audience need to take?
- **What is at stake?:**
  - What are the benefits if your audience acts in the way you want them to?
  - What are the risks if they do not?
- **Form your big idea.**
- It should a) Articulate your point of view, b) Convey what's at stake and c) be a complete and single sentence.

Template from (Knaflic, 2015)

Exploratory visualisation

- Exploration gives an idea about the data we will be digging deep into while analyzing.
- Visualization helps to infer insights easily from massive datasets.
- Spot patterns and anomalies
- Discover trends
- Test hypotheses with summary statistics and visualizations.

All about data

- Before you begin, source data you need.
- This may come from a single source, or from combined sources.
- WARNING:
  - If you are combining sources, make sure they are compatible in their
    - Meaning of observations
    - Relationship (1:1, 1:many, many:many? Matching, left / right join?)

- accumulate data
  - Understand the structure and content of data you have
  - Understand which data you do not have, but would like
  - Tidy the data
- Explore the data
- Formulate the idea you want to share.

Discovering and shaping data

Data completeness

Correct joins

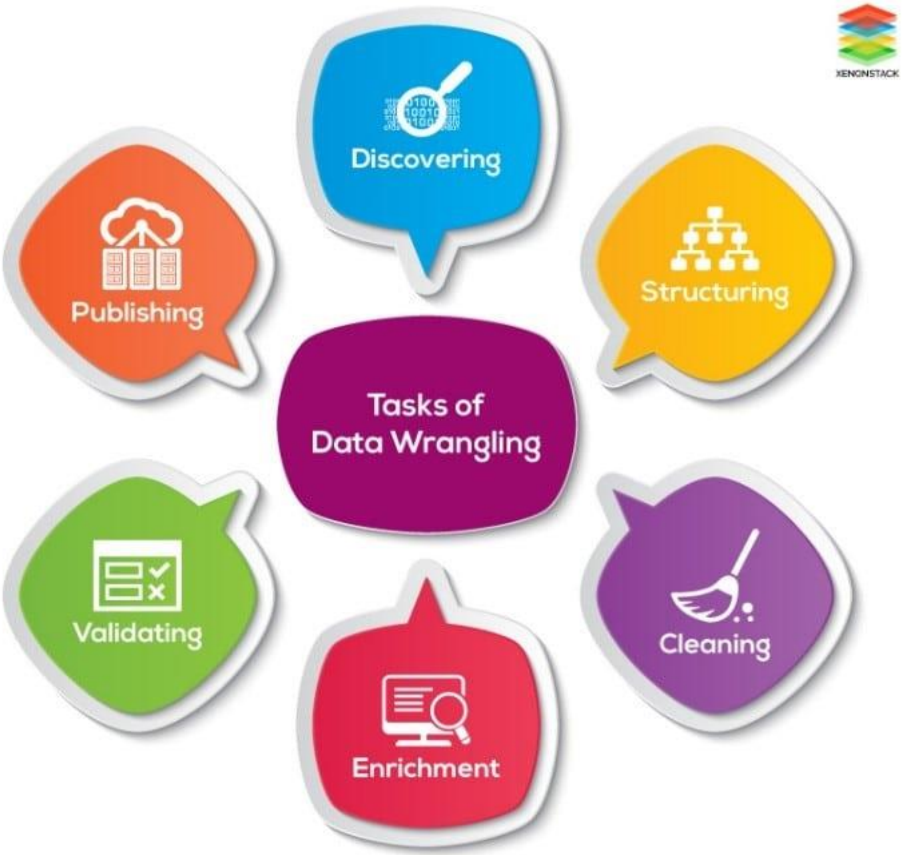
Appropriate formats for data

Datasets downloaded

- Often downloaded as .csv or .json
  - If .json, we can unwind to get single rows.
- Usually not normalized
- May not have any validation
- May be difficult to cross-compare.
- Generally need to be ‘wrangled’

Data Wrangling

1. Discover
2. Structure
3. Clean
4. Enrich
5. Validate
6. Publish



Discovering

- Understand your data
- What is each attribute? ***describe***
- What do they mean? ***Select distinct…***
- Is there a primary key? ***Select … from … group by having count(*)>1***
- How many possible values are there? ***Select count(distinct…)***
- Can it be empty? ***Count… where … is null***
- What is the distribution of the data?
  - Count..group by
- Check statistics in relation to the whole dataset!!!

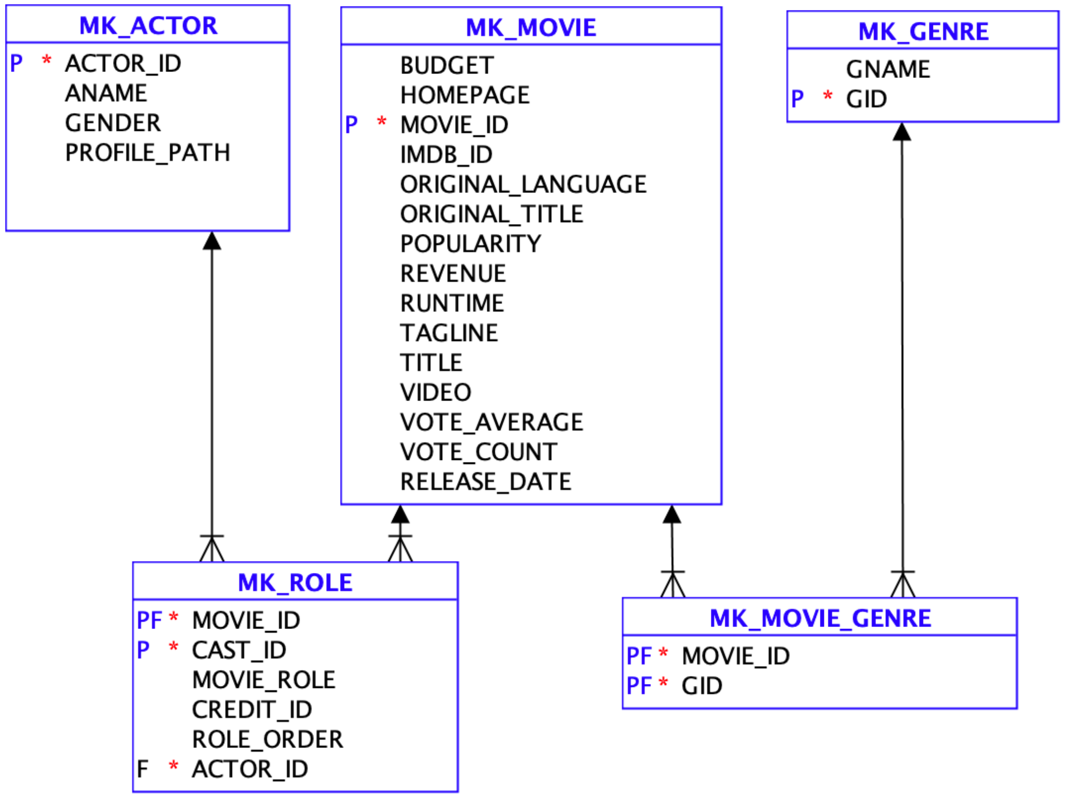
Example supermarket transactions

- A fictitious supermarket logged over 14,000 transactions from across all of its international outlets.
  - Supermarket_transactions.csv

- There is a program that reads the csv and adds it to a SQLite database and runs queries on it.

Structuring

- Data can come in different forms:
  - CSV, XLSX, JSON, DB, DAT, XML
- Investigate relationships within your data
- Restructure your data in the way your application needs it:
  - Relational – normalize
  - Non-relational – denormalize and / or redesign
    - 1:few, 1:many, 1: squillions, many:many



MK_ACTOR
P * ACTOR_ID
ANAME
GENDER
PROFILE_PATH

MK_MOVIE
BUDGET
HOMEPAGE
P * MOVIE_ID
IMDB_ID
ORIGINAL_LANGUAGE
ORIGINAL_TITLE
POPULARITY
REVENUE
RUNTIME
TAGLINE
TITLE
VIDEO
VOTE_AVERAGE
VOTE_COUNT
RELEASE_DATE

MK_GENRE
GNAME
P * GID

MK_ROLE
PF * MOVIE_ID
P * CAST_ID
MOVIE_ROLE
CREDIT_ID
ROLE_ORDER
F * ACTOR_ID

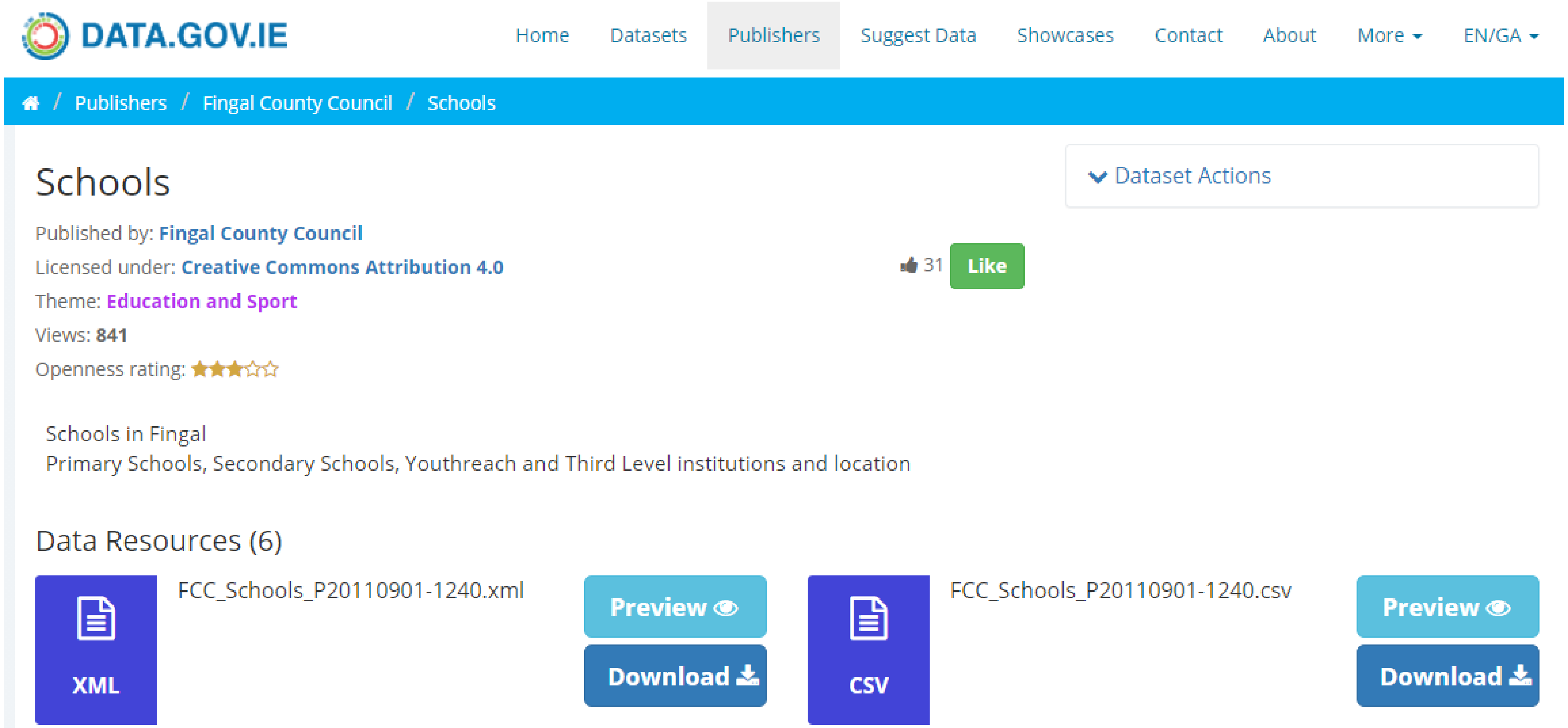
MK_MOVIE_GENRE
PF * MOVIE_ID
PF * GID

Cleaning

- Are there empty values?
- Are there any easily fixed misspellings?
- Are there any really obvious outliers?

- How should you handle them?

Cleaning requirement example



DATA.GOV.IE

Publishers / Fingal County Council / Schools

Schools

Published by: **Fingal County Council**

Licensed under: **Creative Commons Attribution 4.0**

Theme: **Education and Sport**

Views: **841**

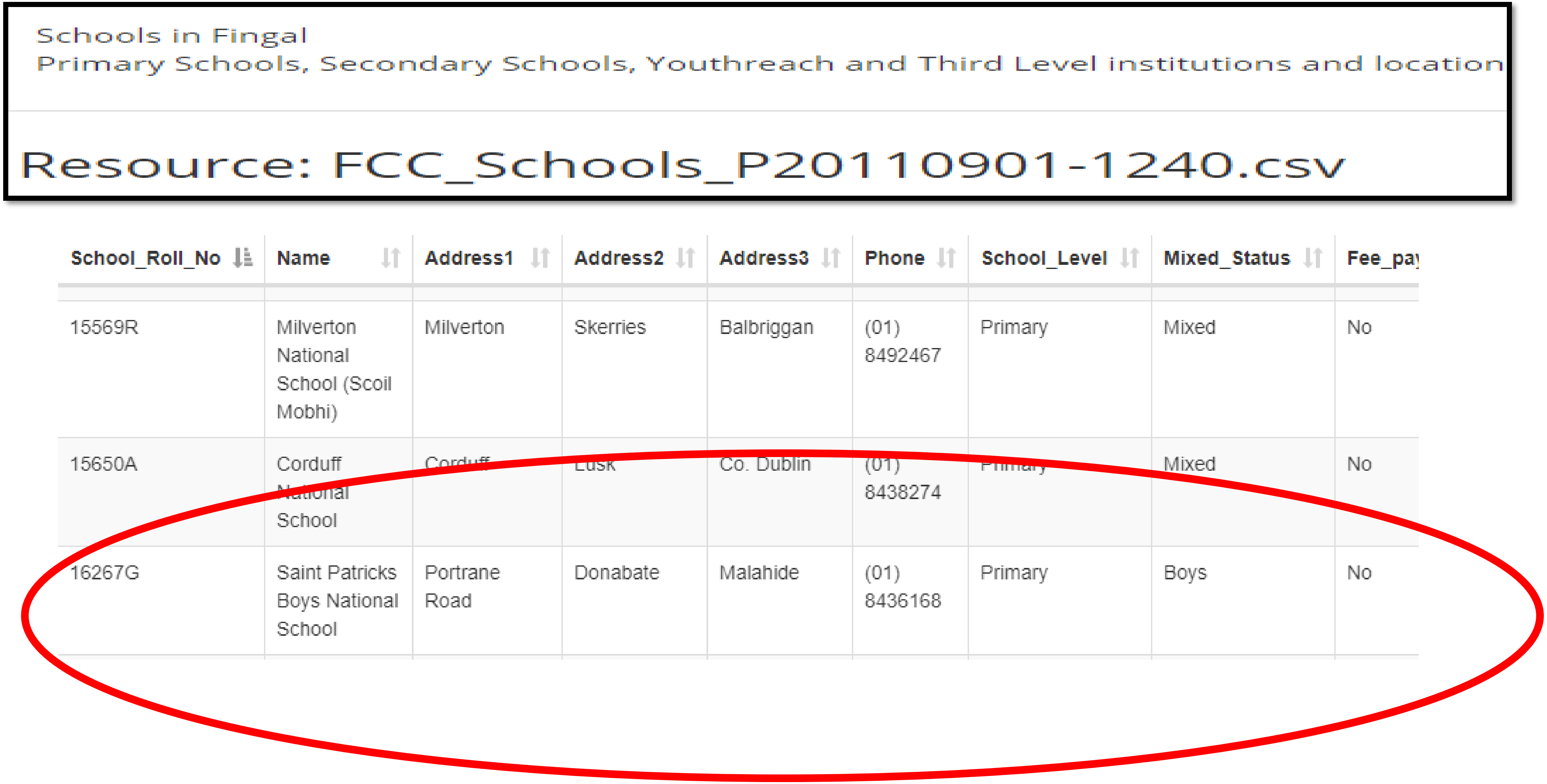
Openness rating: ★★★☆☆

Schools in Fingal
Primary Schools, Secondary Schools, Youthreach and Third Level institutions and location

Data Resources (6)

XML FCC_Schools_P20110901-1240.xml

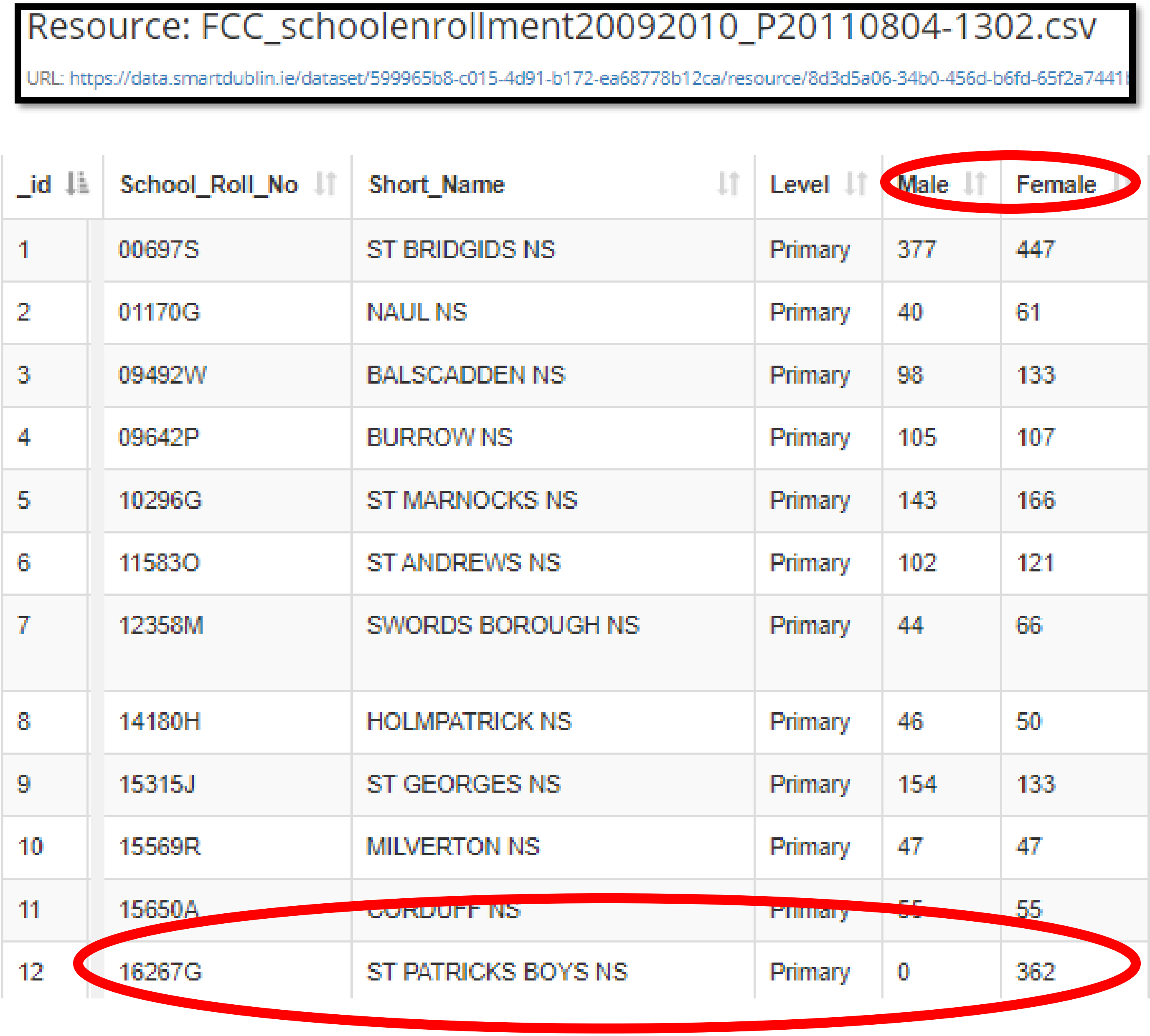
CSV FCC_Schools_P20110901-1240.csv



Schools in Fingal
Primary Schools, Secondary Schools, Youthreach and Third Level institutions and location

Resource: FCC_Schools_P20110901-1240.csv

School_Roll_No	Name	Address1	Address2	Address3	Phone	School_Level	Mixed_Status	Fee_pay
15569R	Milverton National School (Scoil Mobhi)	Milverton	Skerries	Balbriggan	(01) 8492467	Primary	Mixed	No
15650A	Corduff National School	Corduff	Lusk	Co. Dublin	(01) 8438274	Primary	Mixed	No
16267G	Saint Patricks Boys National School	Portrane Road	Donabate	Malahide	(01) 8436168	Primary	Boys	No



Resource: FCC_schoolenrollment20092010_P20110804-1302.csv

URL: https://data.smartdublin.ie/dataset/599965b8-c015-4d91-b172-ea68778b12ca/resource/8d3d5a06-34b0-456d-b6fd-65f2a7441

_id	School_Roll_No	Short_Name	Level	Male	Female
1	00697S	ST BRIDGIDS NS	Primary	377	447
2	01170G	NAUL NS	Primary	40	61
3	09492W	BALSCADDEN NS	Primary	98	133
4	09642P	BURROW NS	Primary	105	107
5	10296G	ST MARNOCKS NS	Primary	143	166
6	11583O	ST ANDREWS NS	Primary	102	121
7	12358M	SWORDS BOROUGH NS	Primary	44	66
8	14180H	HOLMPATRICK NS	Primary	46	50
9	15315J	ST GEORGES NS	Primary	154	133
10	15569R	MILVERTON NS	Primary	47	47
11	15650A	CORDUFF NS	Primary	55	55
12	16267G	ST PATRICKS BOYS NS	Primary	0	362

This dataset comes from the same source as the one on the previous slide.

The school names are different. In at least one case, the values for Male and Female are transposed.

Which data is correct???

Enriching

- Every piece of data you use should be relevant
    - You can drop attributes that aren’t part of your analysis
- You should use every piece of relevant data
    - Don’t drop them if the meaning changes without them.
- Get extra data
    - Join with other datasets, but MAKE SURE YOU’RE JOINING PROPERLY!!!

Validating.

- In SQL, there are constraints
- In MongoDB there are validators
- What are the rules of your data?
- Is one or more row breaking the ‘rule’? What does this mean?

- Have you enriched your data?
- Does the resulting dataset make sense?

Publishing.

- When you are ready to use your data:
  - Cite ALL sources
  - State ALL wrangling steps that have changed the data
  - Describe your attributes, their rules and their associations.

Which tool should I use?

Pandas, SQL, MongoDB?

SQL

- SQL is familiar and can be used in most cases. However, it does have some rules:
  - Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*, 1–23. https://doi.org/10.18637/jss.v059.i10

Tidy data

- Most statistical datasets are
  - rectangular tables made up of
  - rows and
  - columns.

- The columns are almost always labelled and the rows are sometimes labelled.

Tidy data

- A dataset is a collection of values, usually either numbers (if quantitative) or strings (if qualitative).
- Values are organised in two ways.
- Every value belongs to a variable and an observation.
    - A variable contains all values that measure the same underlying attribute (like height, temperature, duration) across units.
    - An observation contains all values measured on the same unit (like a person, or a day, or a race) across attributes.

Tidy data

- Tidy data is a standard way of mapping the meaning of a dataset to its structure.
- A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types.
- In tidy data:
    - 1. Each variable forms a column.
    - 2. Each observation forms a row.
    - 3. Each type of observational unit forms a table

Messy dataset characteristics

- Column headers are values, not variable names.
- Multiple variables are stored in one column.
- Variables are stored in both rows and columns.
- Multiple types of observational units are stored in the same table.
- A single observational unit is stored in multiple tables.