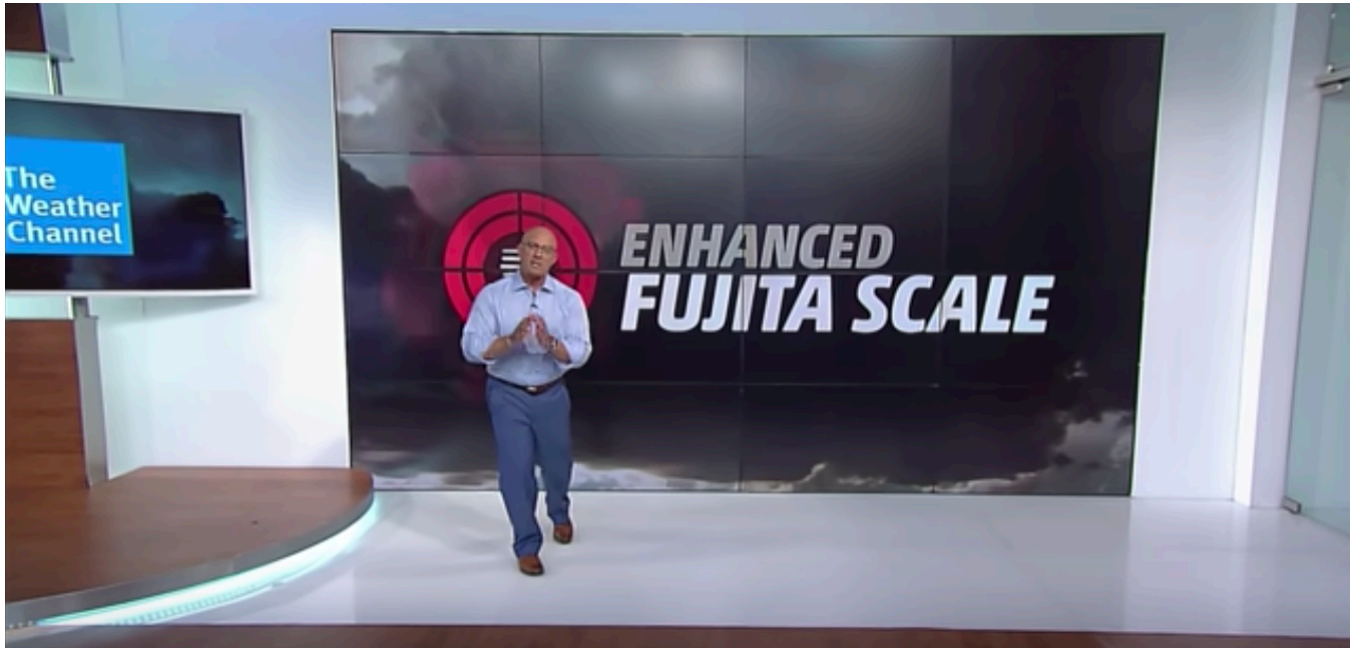
FIT5147 Data Exploration & Visualisation

Kim Marriott

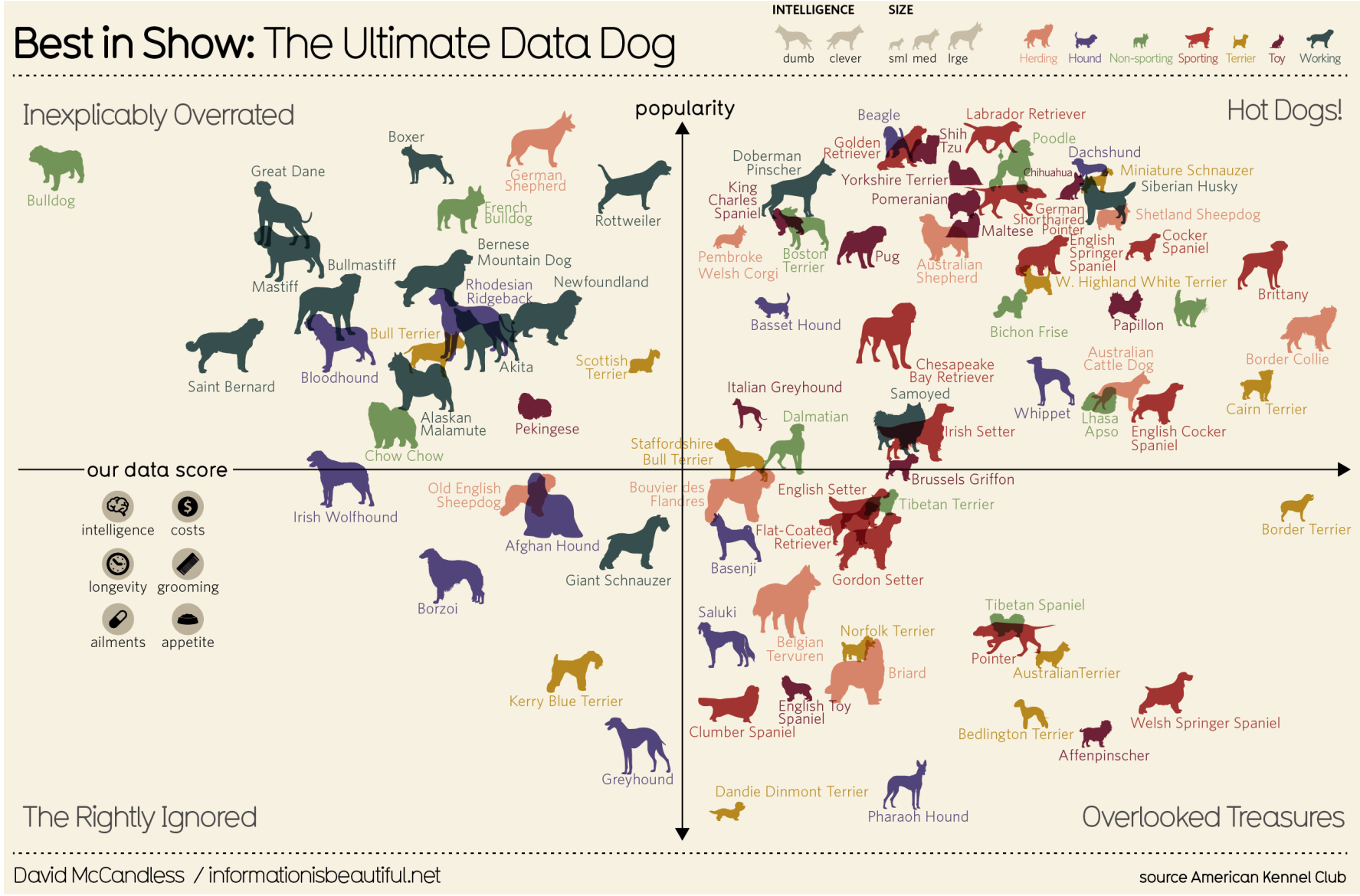
Hall of Fame (Tue 2pm)



https://www.youtube.com/watch?v=0cODBQqaGTw&list=PLki90Aw2GjdeFFwqlQXOaMy6UKihs0TUc /

Arihant Jain

Hall of Fame (Wed 4pm)



Anita Rohmawati

Hall of Fame/Shame

https://coffeemeetsbagel.com/blog/index.php/dating-statistics/7-popular-time-year-relationships-break/

Syllabus

Week	Lecture material	Lab/Tute
. 1	Visual analytics; Tools for data exploration	Intro to Tableau; R; D3
. 2	Visualisation of tabular data	Advanced graphics with R
. 3	Analysis of trends & patterns in tabular data	Interactive graphics with R
. 4	Data maps; Tools for creating data maps	Data maps with R
. 5	Spatial analytics	MapBox; Data Exploration Project feedback
. 6	Network data analysis & visualisation	Relational data and text and text analytics with R
. 7	Textual data analysis & visualisation	Data Exploration Project Feedback
Break		
. 8	Visualisation design methodology	Five design sheet visualisation design methodology
. 9	Human visual system	Introduction to D3
. 10	Visual communication	More D3;Data VisProject Feedback
. 11	Interactive data visualisation	Data Vis Project Presentations
. 12	History and future of data visualisation	Five design sheet visualisation design methodology

Statistical Graphics

1. Paired bar chart
2. QQ Plot
3. Parallel coordinates
4. Bullet graph
5. Heat map (table)
6. Stem-and-leaf plot
7. Spider chart (aka radar chart)
8. Stacked bar chart
9. Compound bar chart
10. Gantt chart
11. Mosaic plot
12. SPLOM (Scatter plot matrix)
13. Area chart
14. Stream graph
15. Wind rose
16. Tree map
17. Doughnut chart
18. Chernoff faces
19. Small multiples
20. Violin plot
21. Hexbin plot
22. Polar area diagram
23. Bubble chart
24. Horizon chart
25. Connected scatter plot

Multivariate Visualisation

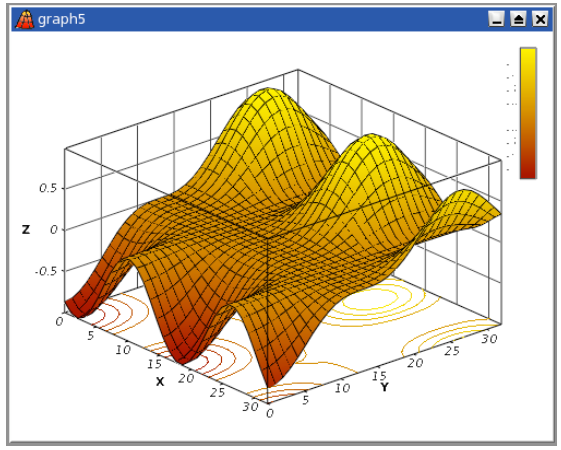
Most real-world data is ***multi-dimensional***

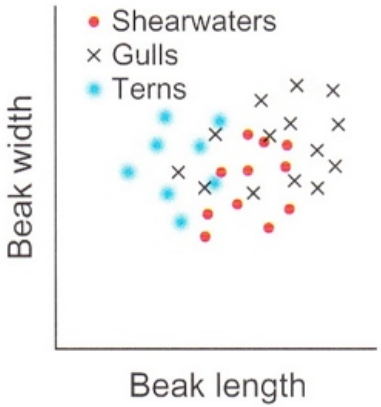
One of the big challenges in data visualisation is dealing with multi-dimensional data (escaping flatland – Tufte)

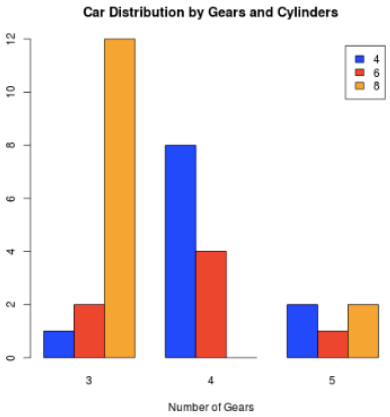
Called **multivariate** visualisation

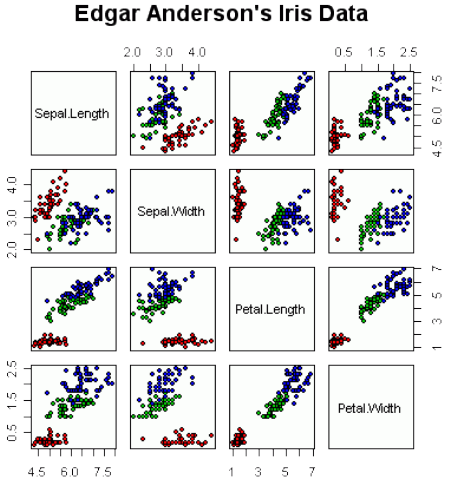
What are some useful multivariate visualisations for multi-dimensional data?

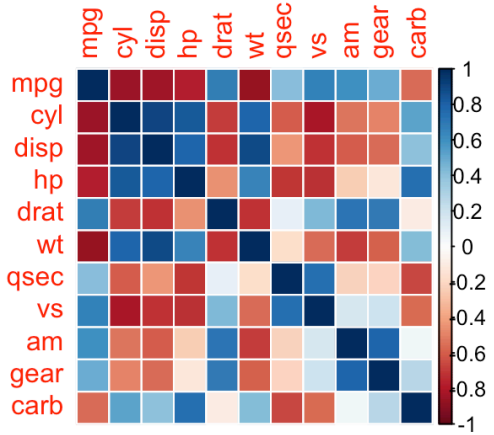
Multivariate Visualisation

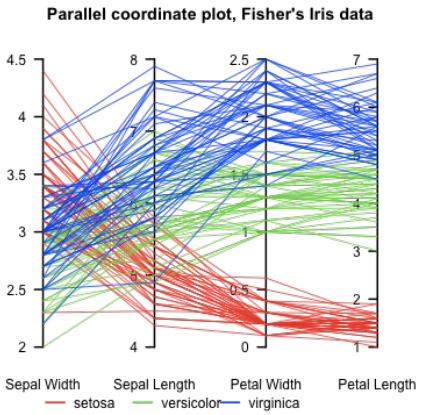


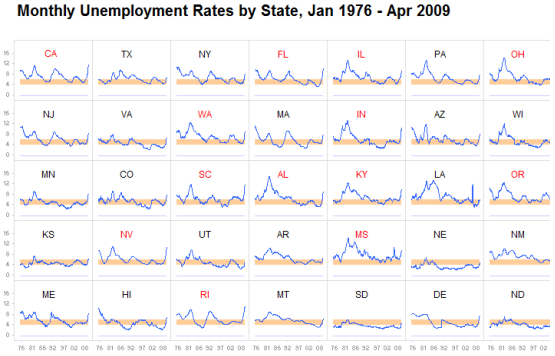












Analytics for Tabular Data

Once the amount of data gets very large we can't visualise all of it

We also want help in drawing conclusions from the data

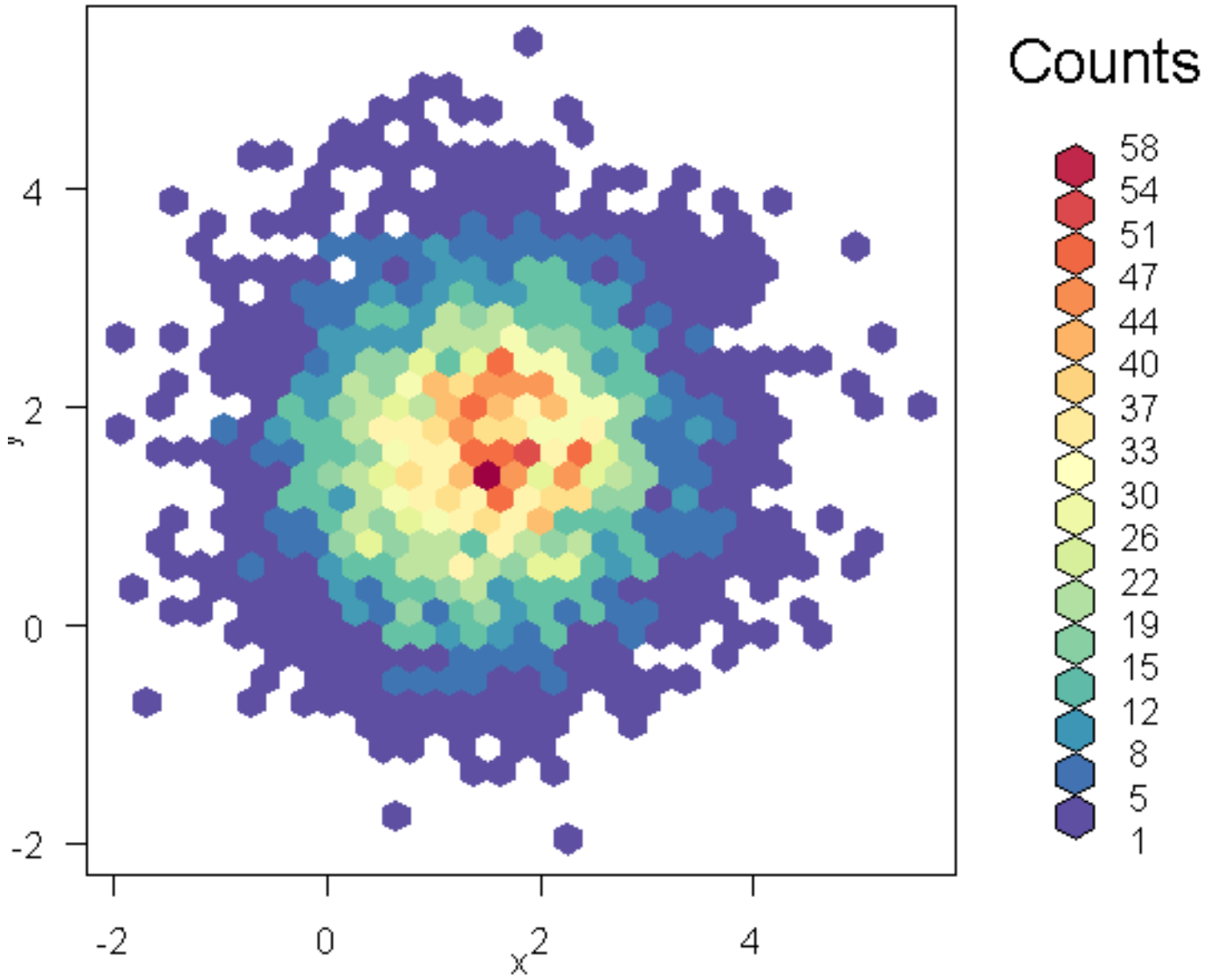
Analytics helps.

What kinds of analytics are there?

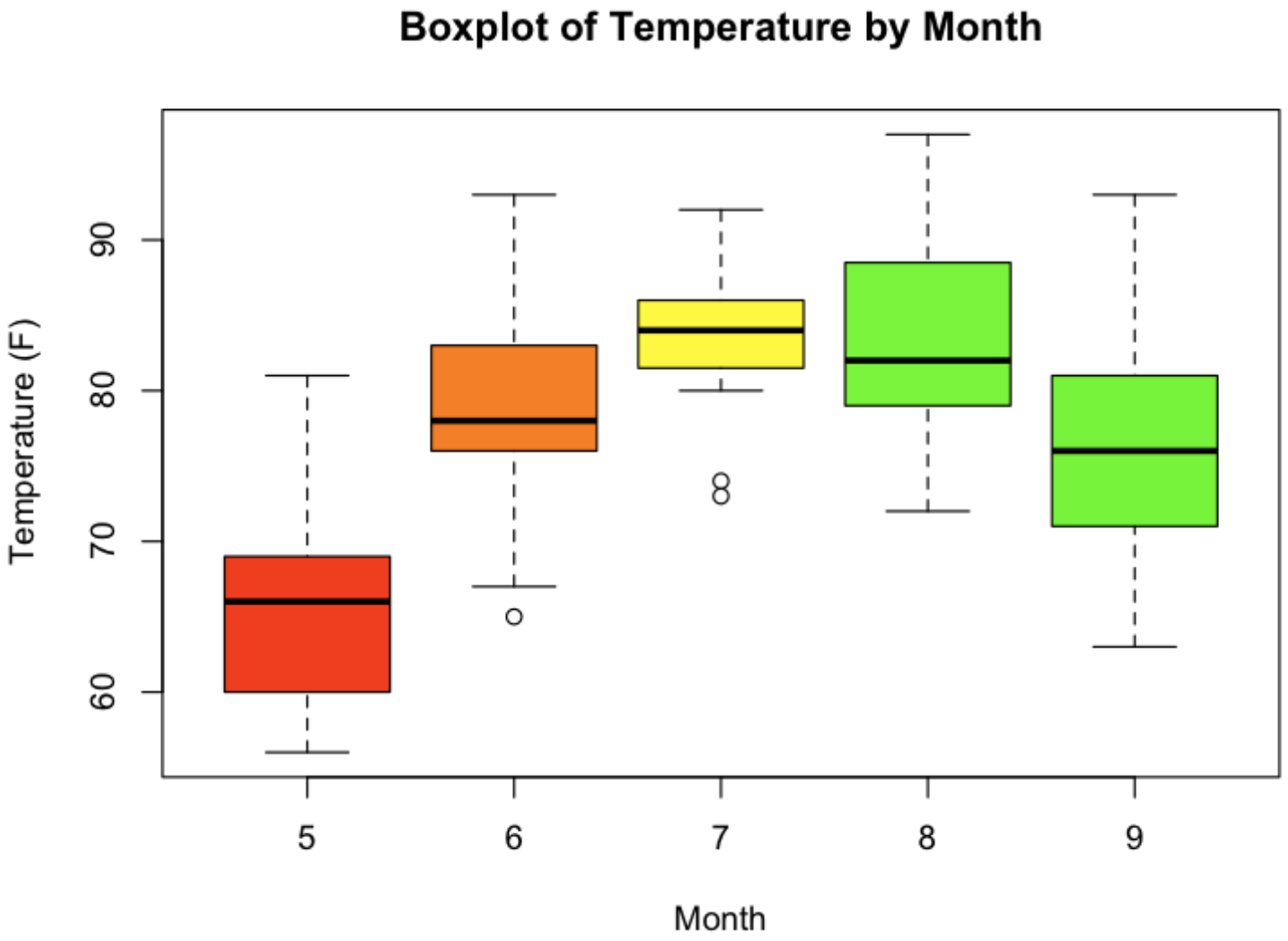
Group Activity

Work in groups of 3-4 to identify 5 ways in which you might explore data using analytic techniques/

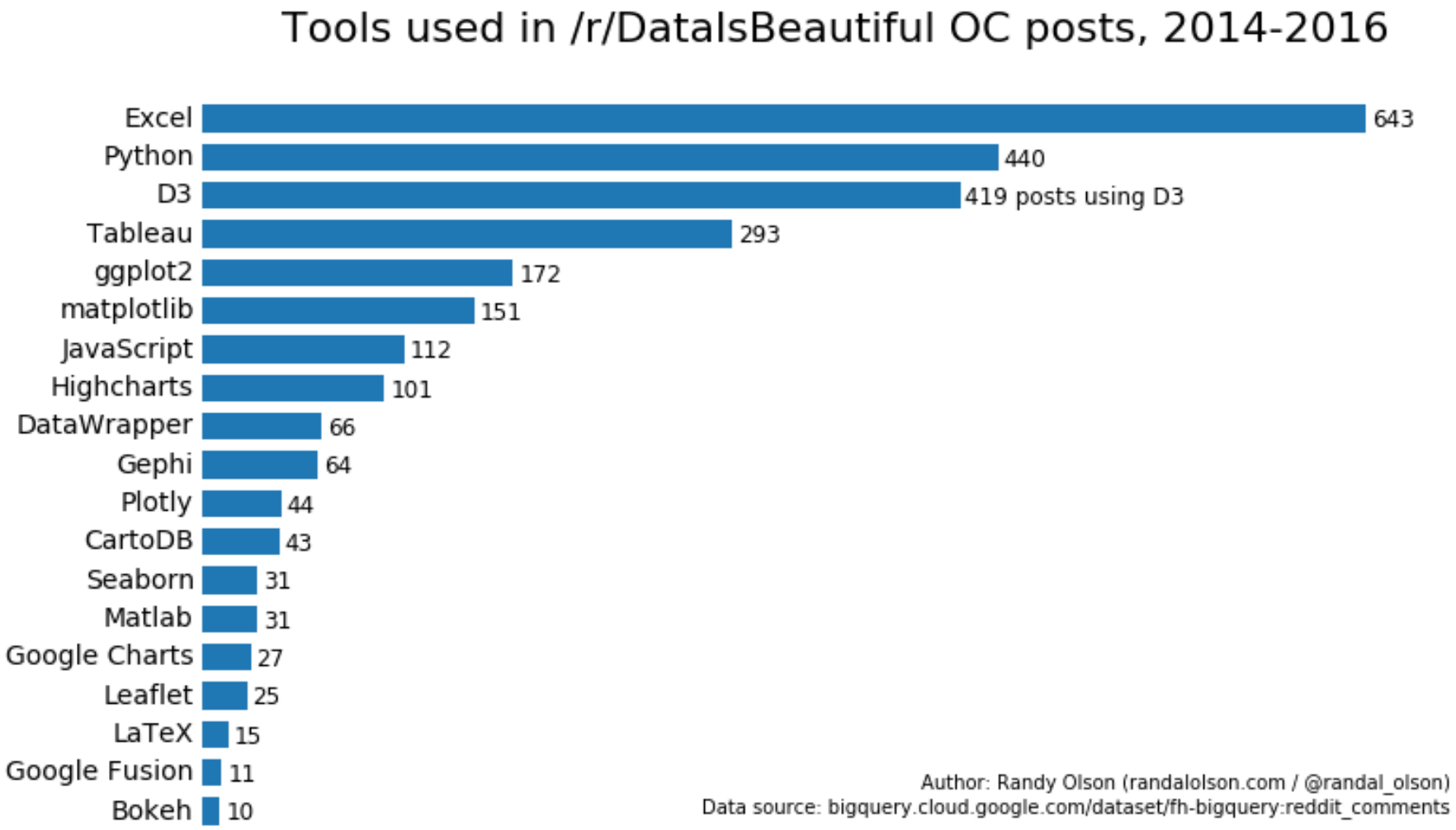
Aggregation



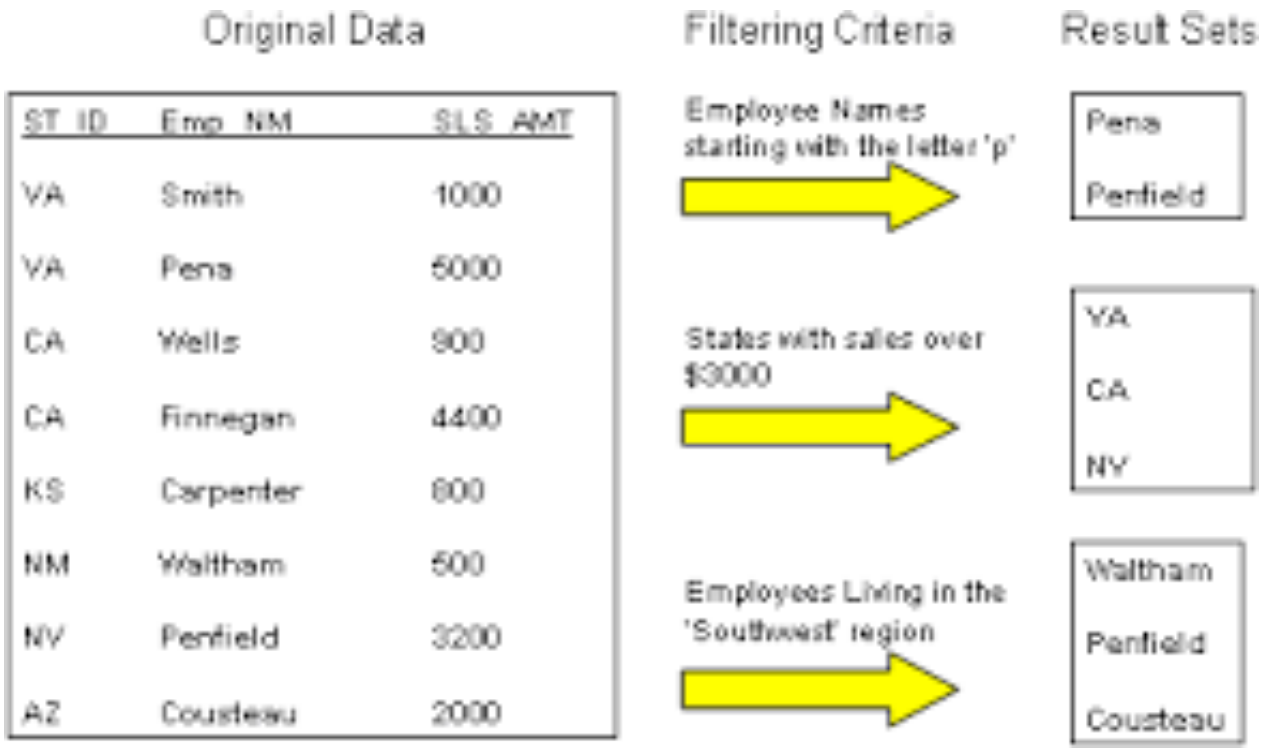
Summary Statistics



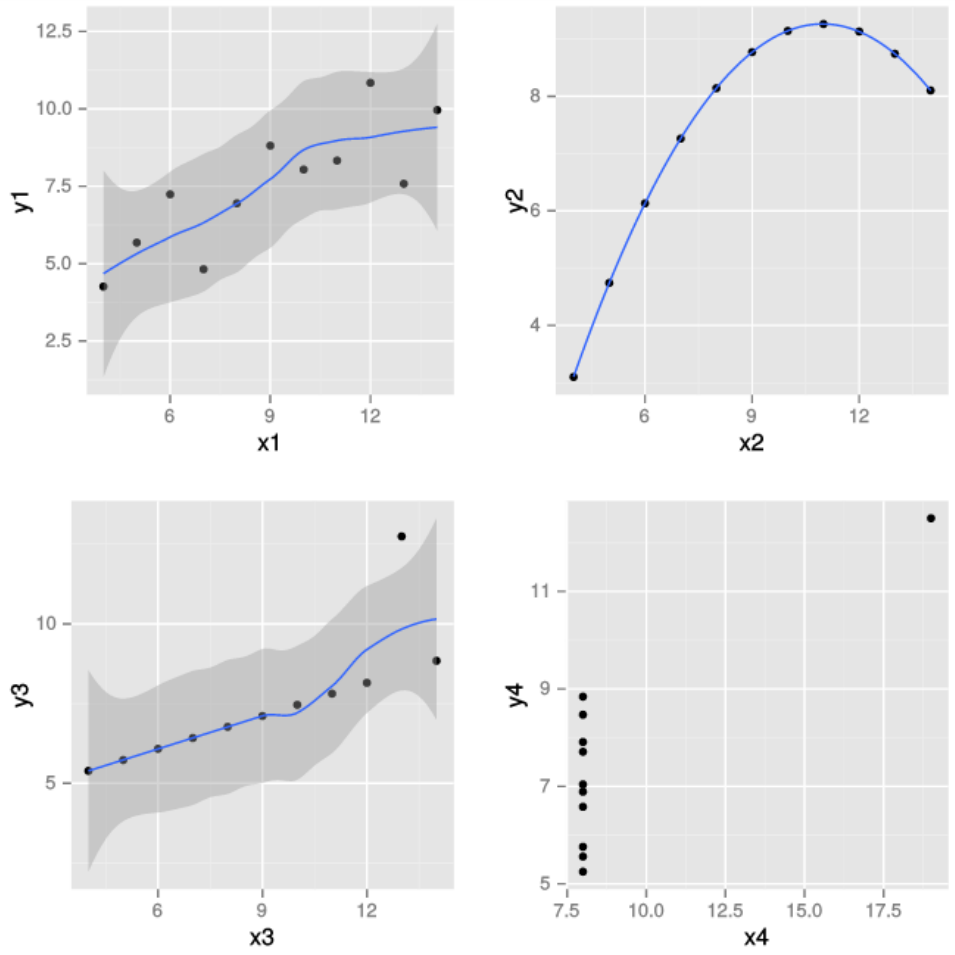
Reordering


Tools used in /r/DataIsBeautiful OC posts, 2014-2016
Excel 643
Python 440
D3 419 posts using D3
Tableau 293
ggplot2 172
matplotlib 151
JavaScript 112
Highcharts 101
DataWrapper 66
Gephi 64
Plotly 44
CartoDB 43
Seaborn 31
Matlab 31
Google Charts 27
Leaflet 25
LaTeX 15
Google Fusion 11
Bokeh 10
Author: Randy Olson (randalolson.com / @randal_olson)
Data source: bigquery.cloud.google.com/dataset/fh-bigquery:reddit_comments

Filtering


Original Data
Filtering Criteria
Result Sets
ST_ID Emp_NM SLS_AMT
VA Smith 1000
VA Pena 5000
CA Wells 900
CA Finnegan 4400
KS Carpenter 800
NM Waltham 500
NV Penfield 3200
AZ Cousteau 2000
Employee Names starting with the letter 'p'
Pena
Penfield
States with sales over $3000
VA
CA
NV
Employees Living in the 'Southwest' region
Waltham
Penfield
Cousteau

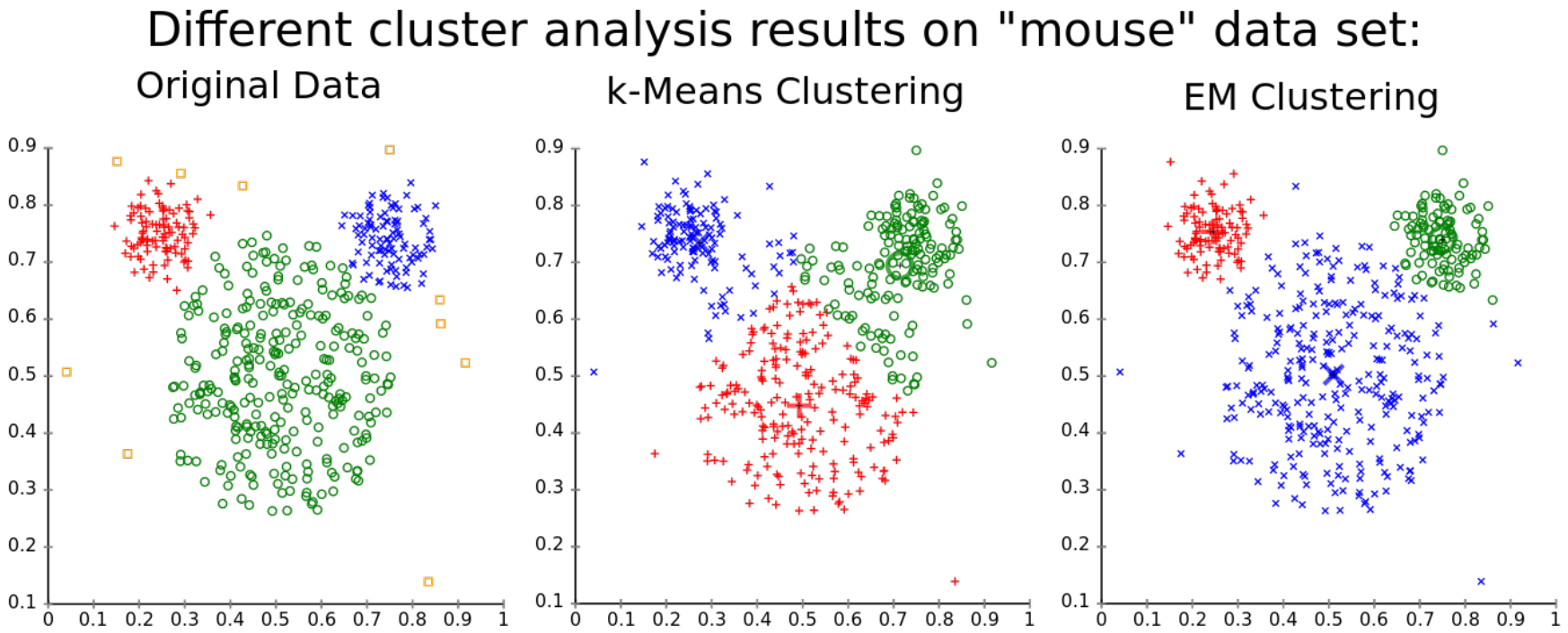
Curve Fitting



Curve fitting is an important part of data exploration.

- Don't just fit a line
- Look at the uncertainty
- Look at the residuals
- Don't just fit a straight line

Clustering

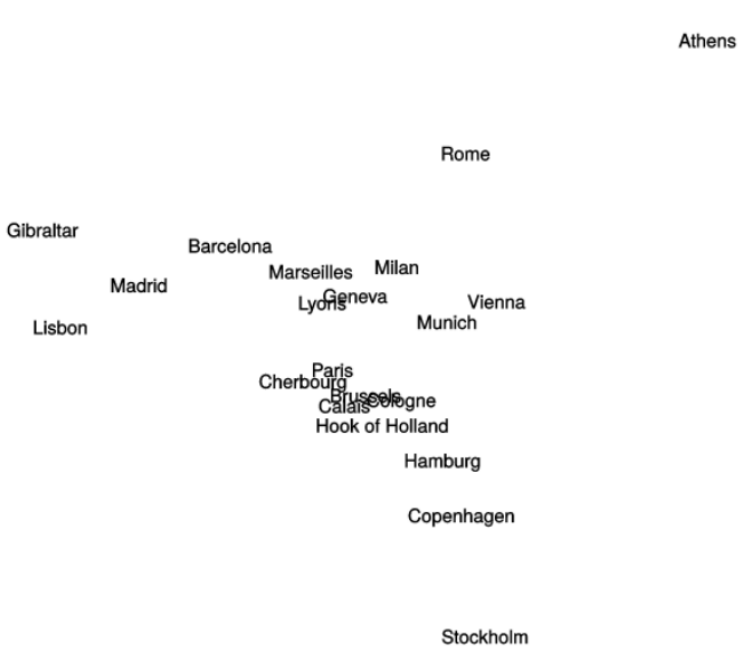


Clustering is another common exploration technique

- Always visually check the results

Dimension Reduction

	Athens	Barcelona	Brussels	Calais	Cherbourg	Cologne	Copenhagen	Geneva	Gibraltar	Hamburg		Lisbon	Lyons	Madrid	Marseilles	Mil
Athens	0	3313	2963	3175	3339	2762	3276	2610	4485	2977	···	4532	2753	3949	2865	22
Barcelona	3313	0	1318	1326	1294	1498	2218	803	1172	2018	···	1305	645	636	521	101
Brussels	2963	1318	0	204	583	206	966	677	2256	597	···	2084	690	1558	1011	92
Calais	3175	1326	204	0	460	409	1136	747	2224	714	···	2052	739	1550	1059	107



This is another technique for handling multidimensional data: Generate attributes for 2D embedding, show these on a scatter plot

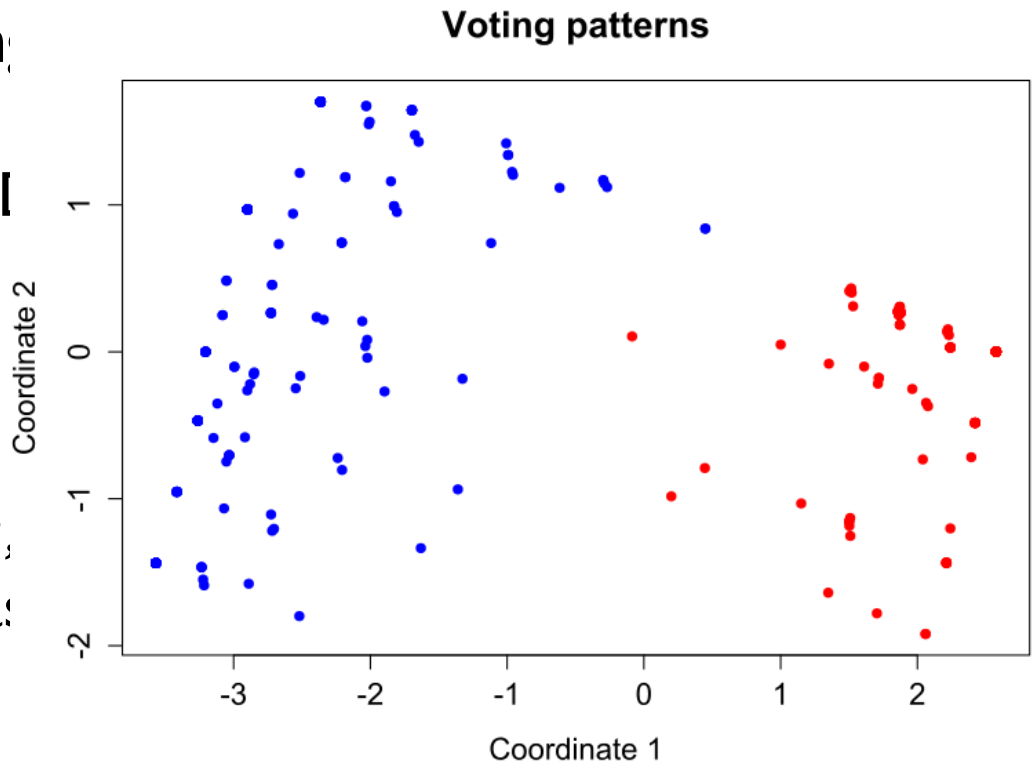
Dimension Reduction

Multi-Dimensional Scalin (MDS) projects a high-dimensional space on to 2 or 3I space.

Data point e is placed at e*

The projection tries to place e_1, so that the distance $||e_1,e_2||$ is a close as possible to $||e_1^*,e_2^*||$

Shows structure of data



Principle Component Analysis (PCA)

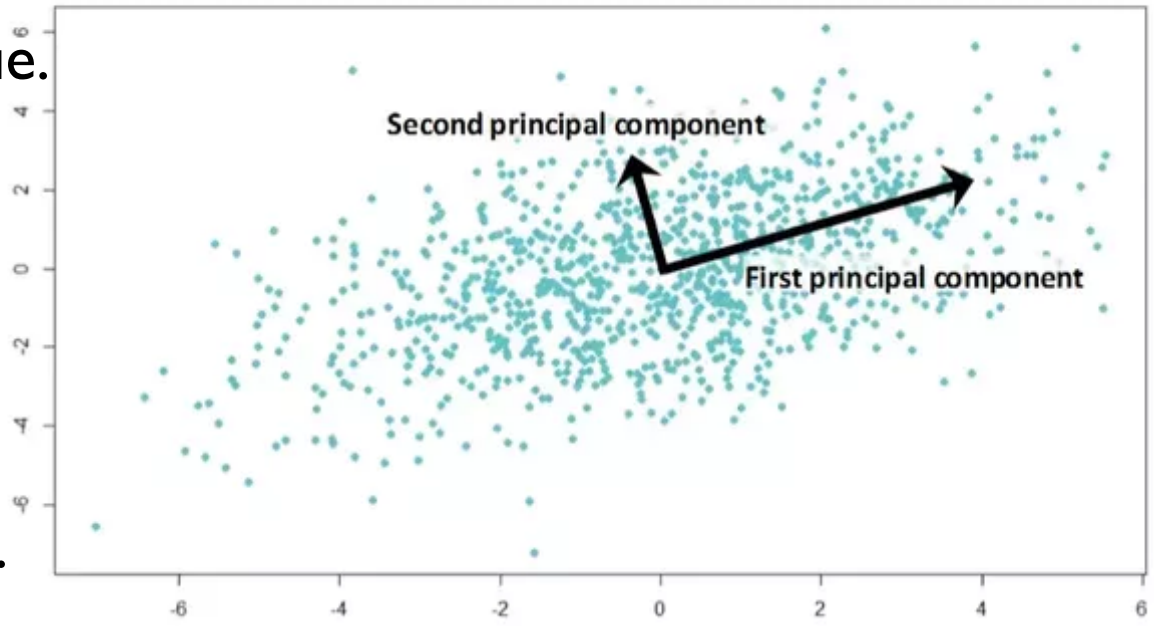
Principle component analysis (PCA) is another dimension reduction technique.

Each axis (called a **principle component**) is a linear combination of the original variables.

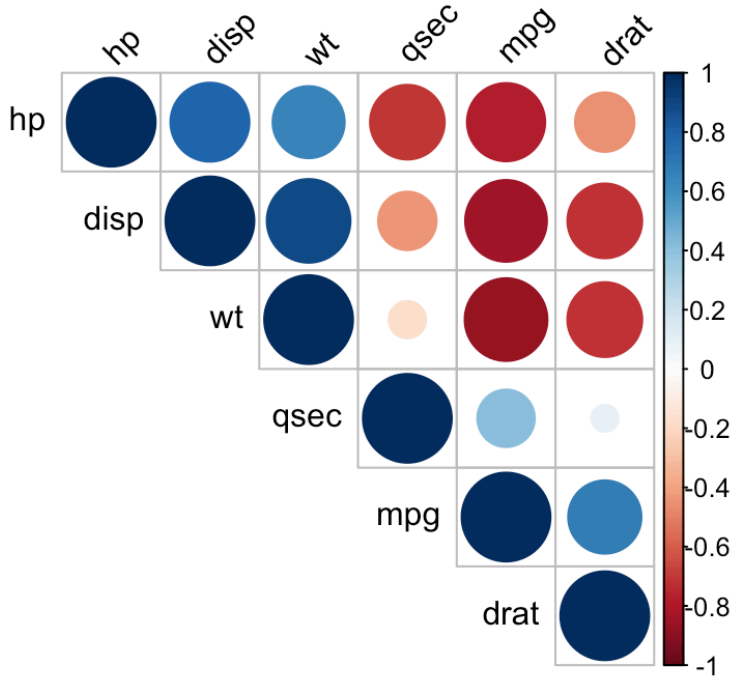
First axis is the line of best fit.

Subsequent axes are orthogonal and maximize remaining variance in the data.

Can be computed really quickly.



Correlation



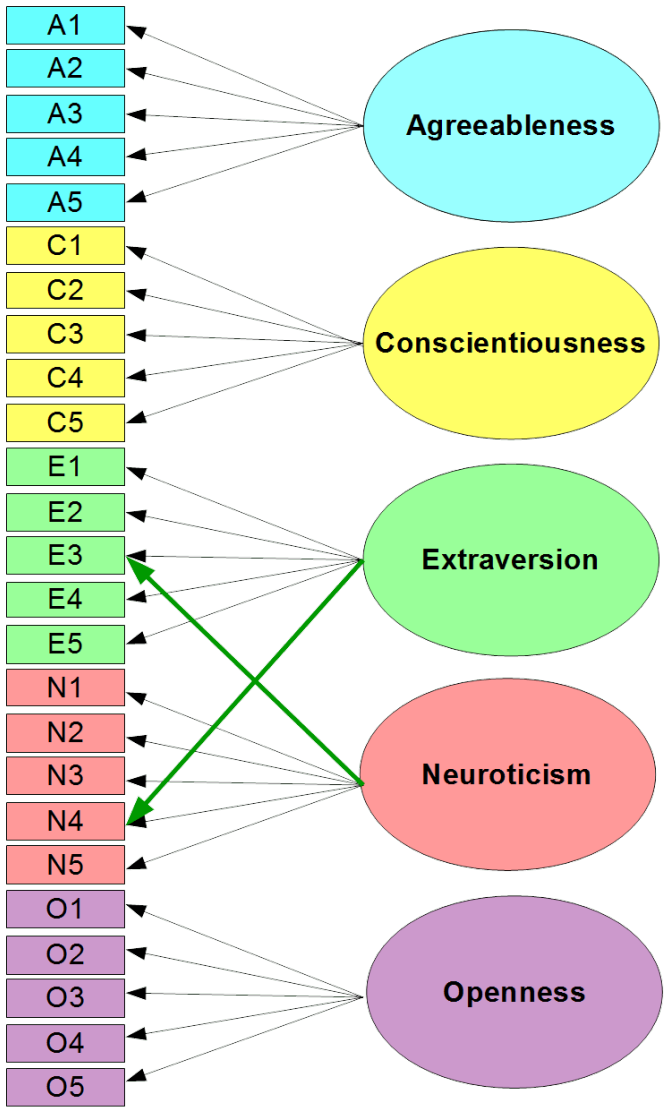
Can compute correlations between all pairs of variables to find which are most highly correlated.

Factor Analysis

Factor analysis extends this by finding hidden or latent variables that explain these correlations.

PCA finds linear combinations of variables that best explain the observations.

Factor analysis finds linear combinations of variables that best explains the relationship between the variables. It clusters the variables rather than the observations.

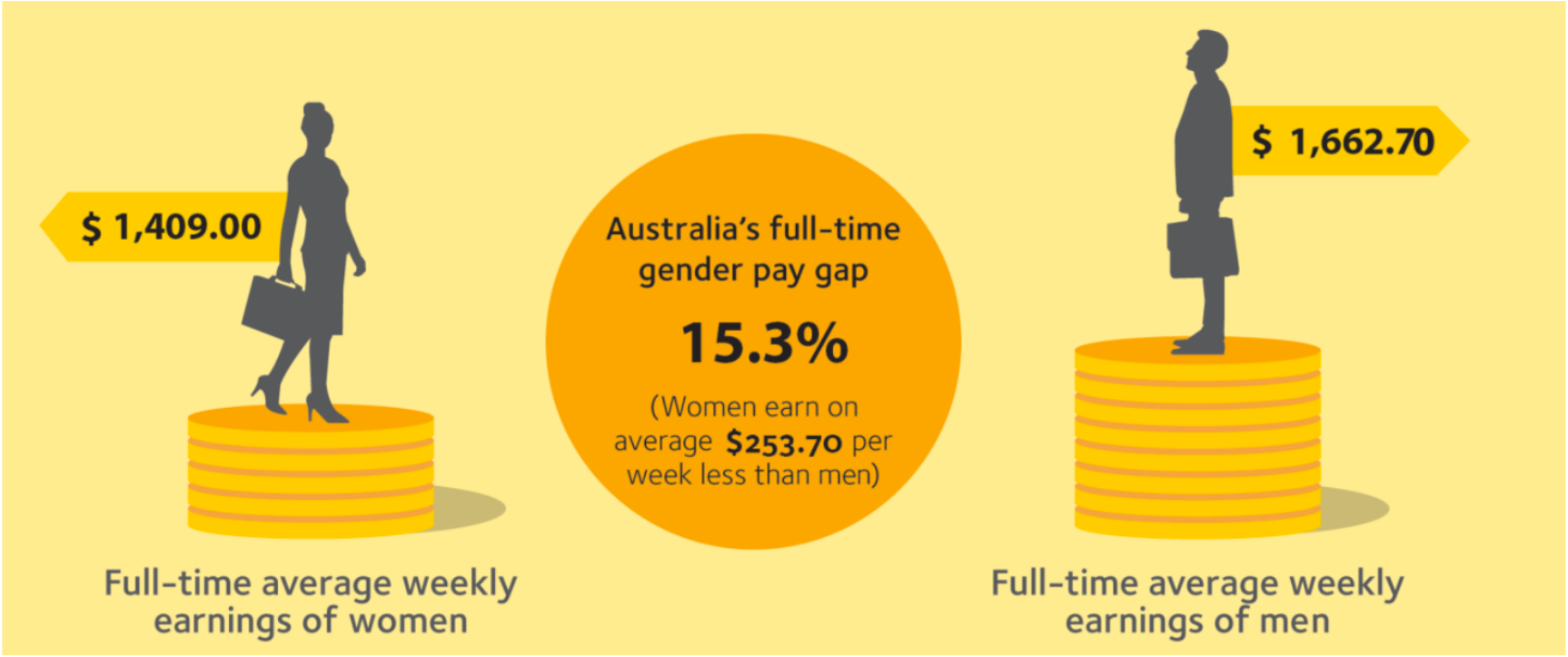


Common Analytics for Tabular Data

- Aggregation
- Summary statistics
- Ranking/ordering
- Filtering
- Curve fitting / Regression
- Clustering
- Dimension reduction
- Correlation
- Factor analysis

February 2018

Australia's gender pay gap statistics



https://www.wgea.gov.au/sites/default/files/gender-pay-gap-statistics.pdf

Group Activity

Work in groups of 3-4 to design visualisations and analyses that you would use to explore the following hypothetical tabular data set in order to understand the wage gap between men and women.

There is an excel spreadsheet for each of the years 2000 to 2016. It has columns

- profession, e.g. pilot, university lecturer
- % of women, men, employed in that profession
- % of part-time, full-time
- % belonging to a union
- minimum and median salary for men, women and overall
- minimum education standard required for the job, median education level

How might you find out if the wage gap is increasing or decreasing?

How might you find out which variables influence the size of the wage gap?

Produce sketches showing what your visualisations might look like and give the kinds of analysis you might use.

Announcements

Data Exploration Project

Once proposal approved by your tutor, get your data into Tableau or R and start exploring it

Online Quiz

Quiz for for week 1-3 materials opens start of Week 4 [details to follow]

Programming Assignment 1: Tableau

Tutors marking, providing feedback in tute, then consultation

Programming Assignment 2: R

- Available end of the week, due week 5
- Cleaned data from Assignment 1
- Create interactive tabular and spatial visualisations using R