Design decisions and dilemmas in a new data science course

David Sterratt

9 May 2022

Outline

- Context
- Intended learning outcomes
- Course description
- Inspiration and courses
- Decisions

Context

- "Inf2 - Foundations of Data Science" (Inf2-FDS, or just FDS) is a new 2nd year undergraduate course, taken by Informatics students on all degree programmes
- 330 students in 2020/21; 290 in 2021/22
- Delivered by Kobi Gal and David Sterratt
- Arises out of the pre-honours curriculum redesign
- Two semester, 20 credit course
- Designed to fit with "Discrete Maths and Probability" (DMP, in S1)
- Runs alongside "Inf2 - Introduction to algorithms and Data Structures"
- Replaces part of "Inf2b - Learning"
- Somewhat less overlap with "Introductory Applied Machine Learning" (year 3 course) than Inf2b had

(Intended) learning outcomes

At the end of the course students will be able to:

1. Describe and apply good practices for storing, manipulating, summarising, and visualising data. (**Data storage, manipulation and visualisation**)
2. Use standard packages and tools for data analysis and describing this analysis, such as Python and LaTeX. (**Tool use**)
3. Apply basic techniques from descriptive and inferential statistics and machine learning; interpret and describe the output from such analyses. (**Stats and ML**)
4. Critically evaluate data-driven methods and claims from case studies, in order to identify and discuss a) potential ethical issues and b) the extent to which stated conclusions are warranted given evidence provided. (**Critical evaluation**)
5. Complete a data science project and write a report describing the question, methods, and results. (**Data science project**)

Course description - technical topics

1. Data wrangling and exploratory data analysis
   - Working with tabular data
   - Descriptive statistics and visualisation
   - Linear regression and correlation
   - Clustering
2. Supervised machine learning
   - Classification
   - More on linear regression; logistic regression
   - Generalization and regularization
3. Statistical inference
   - Randomness, simulation and sampling
   - Confidence intervals, law of large numbers
   - Randomized studies, hypothesis testing

Course description - real-world implications

A. Implications:

- Where does data come from? (Sample bias, data licensing and privacy issues)
- Visualisation: misleading plots, accessible design
- Machine learning: algorithmic bias and discrimination

B. Thinking, working, and writing:

- Claims and evidence: what can we conclude; analysis of errors
- Reproducibility; programming "notebooks" vs modular code
- Scientific communication; structure of a lab report
- Reading and critique of data science articles

Design decision 1: Course outline

- Sketched out by Sharon Goldwater, Heather Yorston & Kobi Gal
- Rough order:
  1. Data wrangling and exploratory data analysis
     - Including linear regression, PCA & K-means
  2. Supervised machine learning
     - Including k-NN and evaluation
  3. Statistical inference (S2)
     - Randomness, sampling, statistical simulations, confidence intervals, hypothesis testing
     - Logistic regression and linear regression using max likelihood
  4. Project

Inspiration & sources

- Berkeley Data 8 - a course introducing programming and statistical inference to students with no programming background
  - Statistical inference via statistical simulations rather than standard probability distributions, as per a standard stats course
  - Python, though using their own `datascience` module
- Devore & Berk *Modern mathematical statistics with applications*
  - Frequentist approach
  - Essentially the same book as used for the DMP course
- Gelman & Nolan *Teaching statistics – a bag of tricks*
- Shannon Vallor's pack *An Introduction to Data Ethics*
- Mine Çetinkaya-Rundel's *Intro to Datascience* course (https://www.introds.org/)

Design decision 2

THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

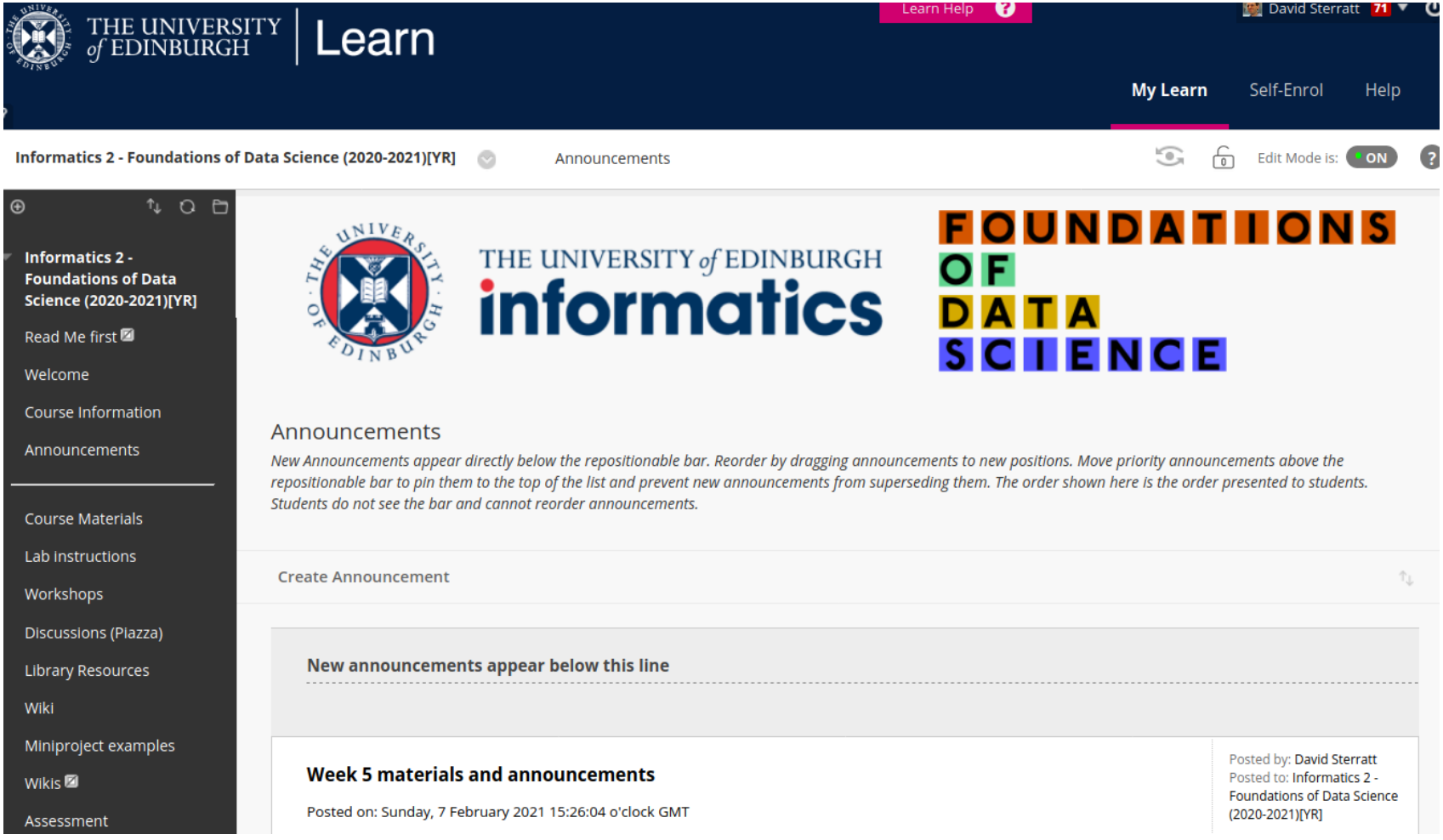
Design decision 3: Assessment

- Originally 10% class test, 30% project, 60% exam
- In July (with other courses) made decision to move to:

	What?		ILOs assessed
20%	Automarked class tests	S1 S2	**Stats and ML**
20%	Visualisation & data manipulation	S1	**Data storage, manipulation and visualisation**, **Tool use**
20%	"Critical evaluation"	S2	**Critical evaluation**
40%	Final project	S2	**Data science project**, **Tool use**, **Stats and ML**, **Data storage, manipulation and visualisation**

- No participation points or microcredit – a mistake?

Design decision 4: Learn to live with Learn



- Why? Familiarity for students & supported tool
- But... Collaborate caused issues for some students -> move to Teams (not popular with all students and staff!)

Design decision 5: Weekly pattern

Statistical inference	
Semester 2, Week 1	• Topic: Randomness, sampling and simulation (00:53:23) • Topic: Estimation: point estimates and confidence intervals (01:31:47) • Coursework: Essay critiquing data science study (academic paper or news article) (20%) [Released Monday] • Reading: *Computational and Inferential Thinking*, Chapter 9 • Reading: *Computational and Inferential Thinking*, Chapter 10 • Reading: *Computational and Inferential Thinking*, Chapter 14 • Reading: *Modern Mathematical Statistics with Applications*, Sections 6.1 and 6.2 • Reading: *Computational and Inferential Thinking*, Chapter 13, *Modern Mathematical Statistics with Applications*, Sections 7.1, 8.1-8.3 and 8.5 • Python Lab: K-nearest neighbours • Workshop: Reflecting on the mini-project • Question and answer session (Moved to next week)
Semester 2, Week 2	• Topic: Hypothesis testing and *p*-values (00:49:30) • Topic: Logistic regression (01:14:40) • Reading: XKCD comic strip on multiple testing - funny! • Reading: *A hypothesis is a liability* - thought-provoking and amusing article • Task: Problem sheet for S2 Week 3 Workshop • Python Lab: Randomness and sampling • Question and answer session (Monday)
Semester 2, Week 3 25-29 January	• Topic: A/B testing (00:35:17) • Coursework Deadline: Essay [Monday] • Reading: *Computational and Inferential Thinking*, Chapter 12 • Reading: *Modern Mathematical Statistics with Applications*, Chapter 10 • Extras • Extra: Bayesian inference applied to A/B testing • Python Lab: Estimation with the bootstrap (including confidence intervals) • Workshop: Problem sheet on S2 Week 1 material • Question and answer session (Monday)

- Important to give students a rhythm
- Took to it after a few weeks

A topic

Topic: Dealing with high dimensions

Learn Help

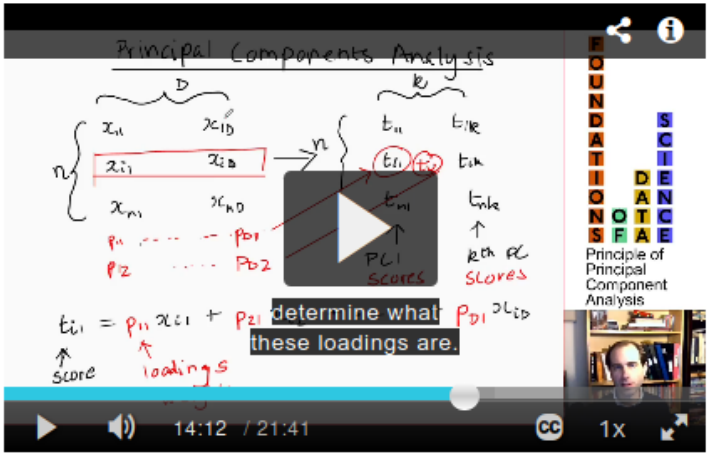
Build Content | Assessments | Tools

Lecture notes and slides

Attached Files:
- FDS-7-1-high-dimensions.pdf (1.07 MB)
- FDS-7-1-1-principle-pca.pdf (12.892 MB)
- FDS-7-1-2-principle-finding-pcs.pdf (635.586 KB)
- FDS-7-1-3-pca-and-regression.pdf (642.957 KB)

The principle of principal components analysis

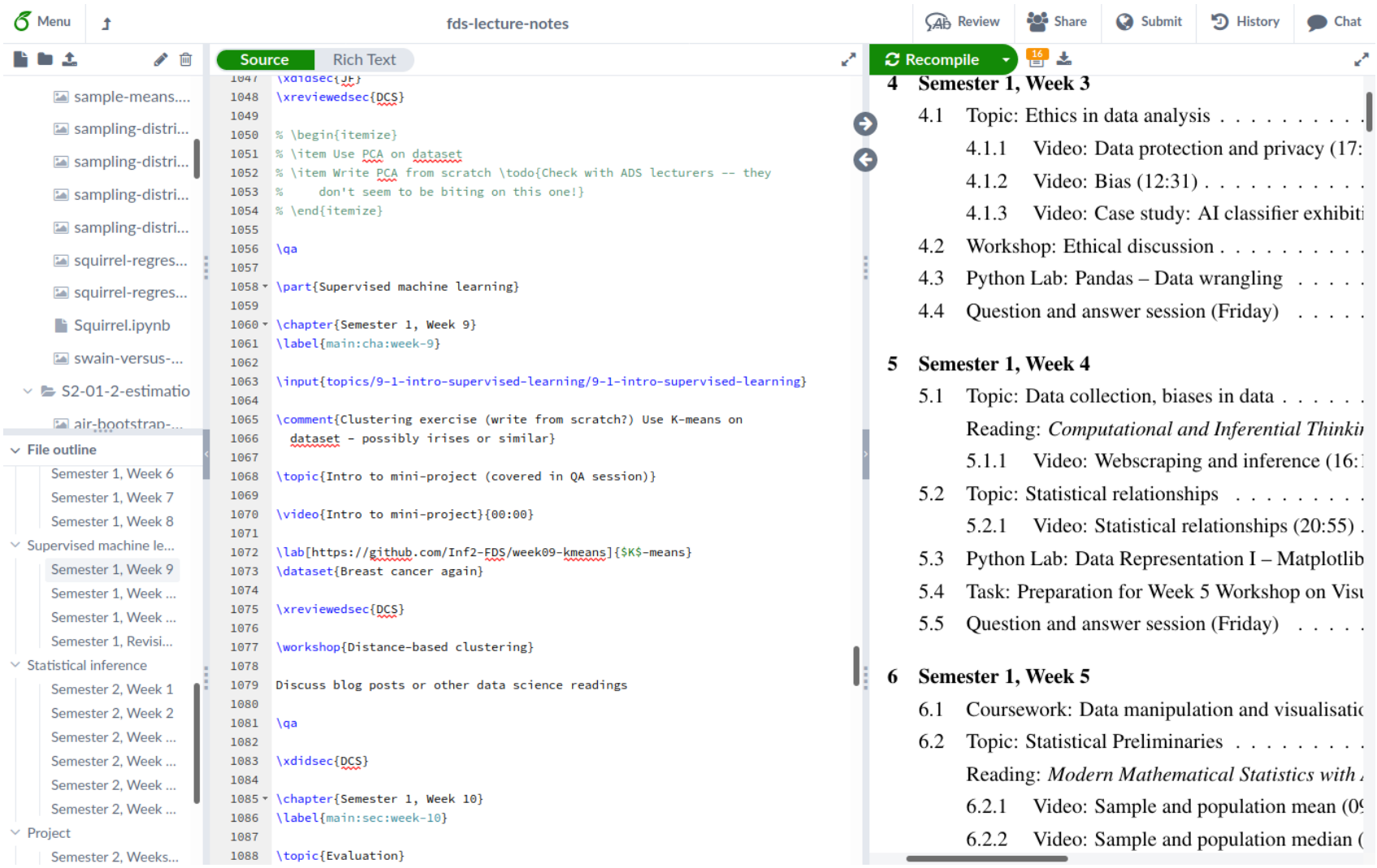
Enabled: Statistics Tracking



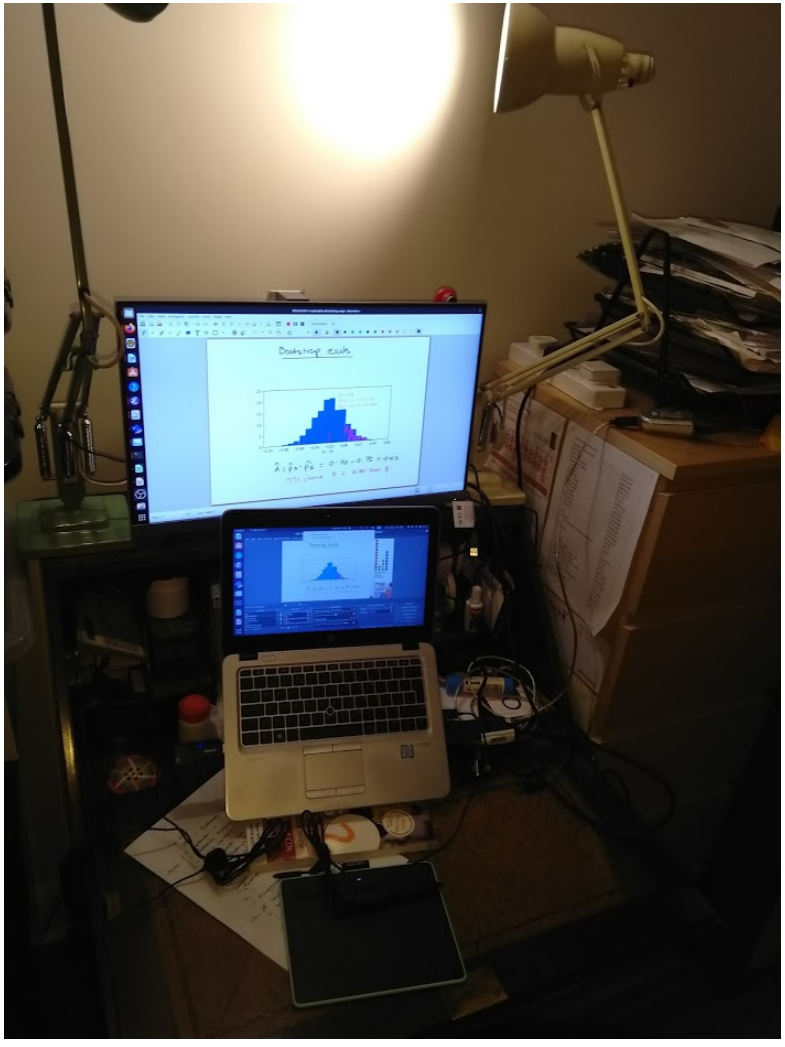
More details

- Video lectures, grouped in 1 or 2 topics
- Most topics have lecture notes attached
- "Comprehension questions" on topic – actually a bit more than comprehension.
  - Purely optional, but popular, especially before class test
- Lab
- Every other week: "Workshop"
- QA session every week - one used for a guest lecture

Design decision 6: Write lecture notes before lecture recording



Design decision 7: Lecture recording



My new best friends

Setup details

- OBS (recording)
- Wacom Tablet
- Xournal++ (whiteboard)
- Kdenlive (editing)
- Media hopper create (hosting and captioning)

Design decision 8: Ethics

- Concept is that ethics is embedded in the course, not just in the final lecture
- Started with group presentation and discussion based around scenarios in Shannon Vallor's pack:
  - Facebook's emotion manipulation experiment (informed consent)
  - OK Cupid data breach (legal, terms and conditions)
- Also mentioned algorithmic bias and transparency and interactions between law and outcomes in credit approval (invited lecture by Dr Galina Andreeva from the Business School)
- Questions on ethics and law in class tests and coursework
- Students seem to have been interested

Design decision 9: Collaboration groups

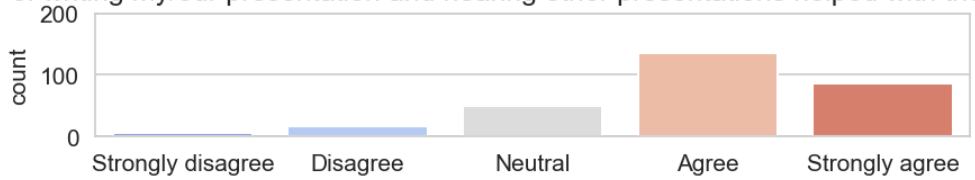
- Workshop $\approx$ Tutorial
- Split workshop groups of 16 into 4 "collaboration groups" of 4, and set tasks
- Worked well for first 2 workshops. . .
- Worked well for some students:
  "Love the group work, I've gotten to know my group well and they're lovely. Its really refreshing to chat to new people"
- . . . But not others:
  "Managing the collaboration group can be tricky"
  "Maybe add a symbolic grade/engagement, to produce a more engaged discussion."

Design decision 10: The final project, Weeks 7-11

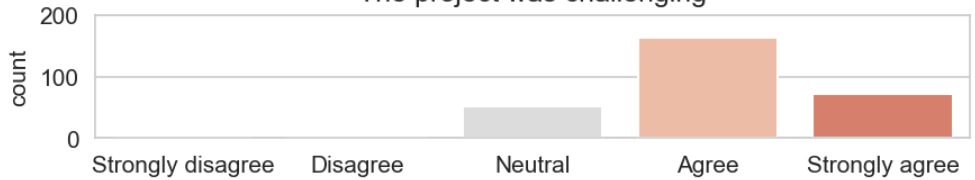
- Three public datasets:
  - 20-21: Edinburgh Just Eat Bike use, Higher Education stats, Scottish Munro features)
  - 21-22: Scottish A&E waiting times, Spotify, EEDI education dataset
- Seed question + ideas for future questions
- Project can be undertaken by individuals, pairs or threes (page limits of 6, 8 and 10 respectively)
- No lectures, but workshops at which lightweight presentations are made. (Not marked, but prerequisite)
- Team formation & code submission via GH Classroom
- Form for assessment of contributions
- Lots of python scripts to administer

Feedback on the project

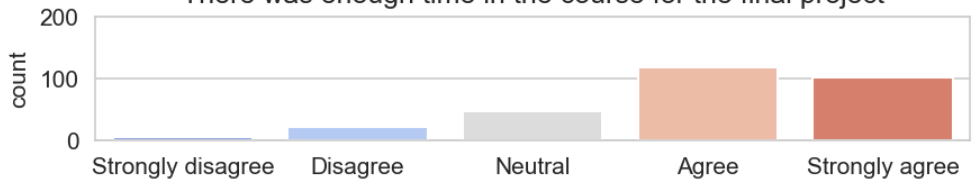
The process of writing my/our presentation and hearing other presentations helped with the final project report



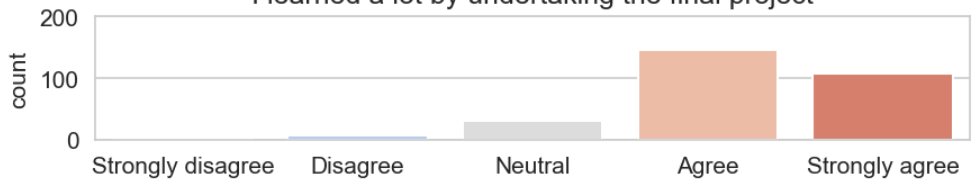
The project was challenging



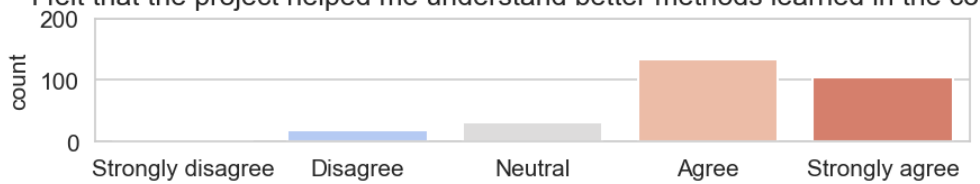
There was enough time in the course for the final project



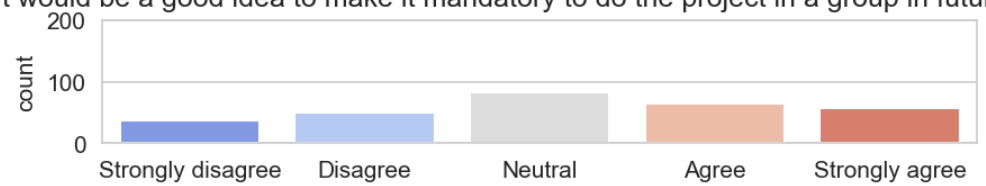
I learned a lot by undertaking the final project



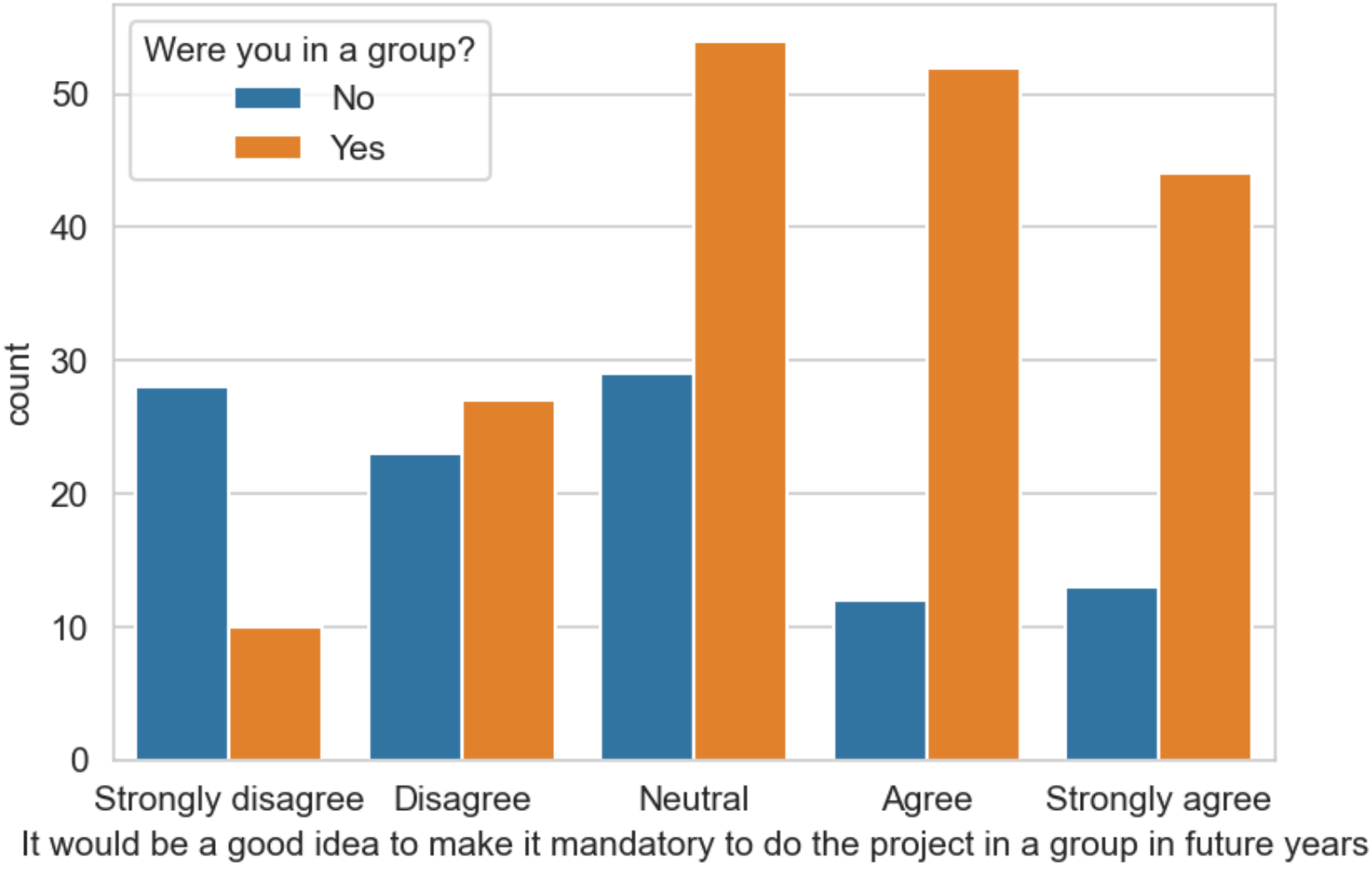
I felt that the project helped me understand better methods learned in the course



It would be a good idea to make it mandatory to do the project in a group in future years



Should groups be mandatory?



Being online

- Videos seem to work well & comprehension questions seem to work well
- Q&As appreciated by those who attend (20-70)
  I think the comprehension quizzes are a great way to see if I actually absorbed the knowledge from the slide and the q&as are great.
- Workshops and labs are harder...

How it started. . .

. . . how it's going

2021/22: Going hybrid

- Labs in person
- Workshops moved to in person, using teaching studios
  - Worked well, allowing ad-hoc collaboration groups to form
- Re-used recorded lectures (with some corrections)
  - Not universally popular...
  - ...will move to in-person
- Assessment structure similar to last year
  - Gradescope used where appropriate
  - Modified rubrics

Rules for visualisation

Informatics 2 – Foundations of Data Science: Visualisation principles and guidance

Principle 1: Show the data

Aim to show as much of the data as possible without leading to a confusing visualisation. There are often multiple ways of representing the same dataset, and no "right" answer. The following guidance on arranging the the graphical elements of the plot should help you to show as much of the data as possible:

- **Choose an appropriate plot type.** Some basic types are:
  - Bar charts: good for plotting numeric variables associated with categorical or ordinal variables, for example the mean weight (numeric variable) of male and female (categorical variable) squirrels.
  - Line charts: good for showing trends of numerical variables over time (a numerical variable).
  - Scatterplots: Scatterplots show the relationship between two numeric variables.
  - Boxplots: Boxplots are a way of representing the distribution of a numeric variable for multiple categories. For example, the median and inter-quartile range of the weights of male and female squirrels.
  - Histograms and density plots: These univariate plots are good for showing the distribution of a single variable.
- **Show multiple variables by using length, shape, size and colour:**
  - The above plots are all bivariate, since they show the relationship between two variables. Use shape and colour to create extra dimensions for categorical variables. For example, in a scatterplot of squirrel weight versus length, we can indicate sex using colour. We could also indicate our age categories by changing the size or the shape of the markers. However, adding information using marker properties can detract from the plot.
  - Barcharts can be extended to two categorical variables and one numerical variables by using colour.
- **Use colour effectively.** (Wexler et al., 2017, pp. 14–18)
  - Choose an appropriate colour scale, depending on if the data is sequential, diverging or categorical.
  - Colour can also be used to highlight features in the plot, e.g. the largest two bars in a bar plot.
- **Encourage the eye to compare several pieces of data**, e.g. by using multiple plots with the same scale.
  - Wexler et al. (2017), p. 31, is a nice example of how this can work better than using multiple symbols on a plot (p. 30).
- **Present many numbers in a small space**
  - A boxplot takes up as much space as a barplot, but conveys more information. For example, a boxplot of the squirrel's weight versus sex shows information about the distribution of the weight as well as the mean weight.
- **Choose appropriate transforms**
  - Sometimes it can make sense to transform data so that features of it are clearer. For example, a time series of Bitcoin over time will show very little detail about the early history of the currency, when it was not valuable. However, plotting the log of the value of Bitcoin on the y-axis allows this detail to be seen.

Principle 2: Make the meaning of the data clear

A visualisation is meaningless if it's not labelled. Every plot should have:

- Title or caption
- Axis labels as English words
- Units given, where appropriate (e.g. "Length (mm)" *not just* "Length")
- All variables labelled – e.g. a legend indicating the colours used to represent squirrel sex
- Use graphical and textual annotation – e.g. it can be helpful to highlight a time series with events that you know about

Principle 3: Avoid distorting what the data have to say

Choices in visualisation design can lead to the instant impression given by preattentive processing of the visualisation being quite different to the numbers in the dataset. Tufte (1982) measures the level of distortion in a visualisation by the "Lie factor":

$$\text{Lie factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

The following guidelines should help to avoid distorting the data:

- **Use appropriate scales and baselines**
  - A very common problem is that the baseline (i.e. the lowest point on the $y-$axis) in a barchart is not zero. This can lead to small differences appearing large.
- **Be aware of limitations of our perception of size**
  - Although marker area can be useful for indicating categories, humans are not very good at relating the area to a quantity – we are much better at comparing lengths.

Principle 4: Make the data accessible

A visualisation is meaningless if it's illegible and loses impact if it's difficult to read. To ensure data is accessible:

- **Make sure text is legible**, i.e. font size of minimum 8 points in a PDF, or about 20 points in a presentation. (It is surprising how often talks are given in which it's impossible to read the labels on plots even from the front row.)
- **Use colours that work for people with colour-vision deficiency.** Wexler et al. (2017), Chapter 1 has an excellent introduction to using colour in visualisations.

Principle 5: Focus on the content

Give the viewer's brain as little work to do as possible.

- No chartjunk – e.g. colours that don't have any meaning.
- Reduce clutter
- Consistent colours between plots in a study
- Correct spelling

References

Tufte, E. (1982). *The visual display of quantitative information*. Graphics Press, Cheshire, Connecticut.

Wexler, S., Shaffer, J., and Cotgreave, A. (2017). *The Big Book of Dashboards: Visualizing Your Data Using Real-World Business Scenarios*. Wiley.

Not discussed

- Rules for visualisation?
- Application versus derivation?
- Lab development

Acknowledgements

- Kobi Gal
- Sharon Goldwater, Heather Yorston
- Chris Williams
- The marvellous team of FDS tutors and lab demonstrators, including Jonathan Feldstein on Lab development
- ILTS
- Other colleagues
- My family