DS Visualisation and Analysis

Abbey Waldron

Today

- Project hand in start
- Exam preparation
- Project presentations start

Final Projects

Please hand it in now - the sooner you hand it in the sooner you can leave.

Final deadline for hand in is 11:20 am today.

Do not leave before you have given your presentation to me!

Exam

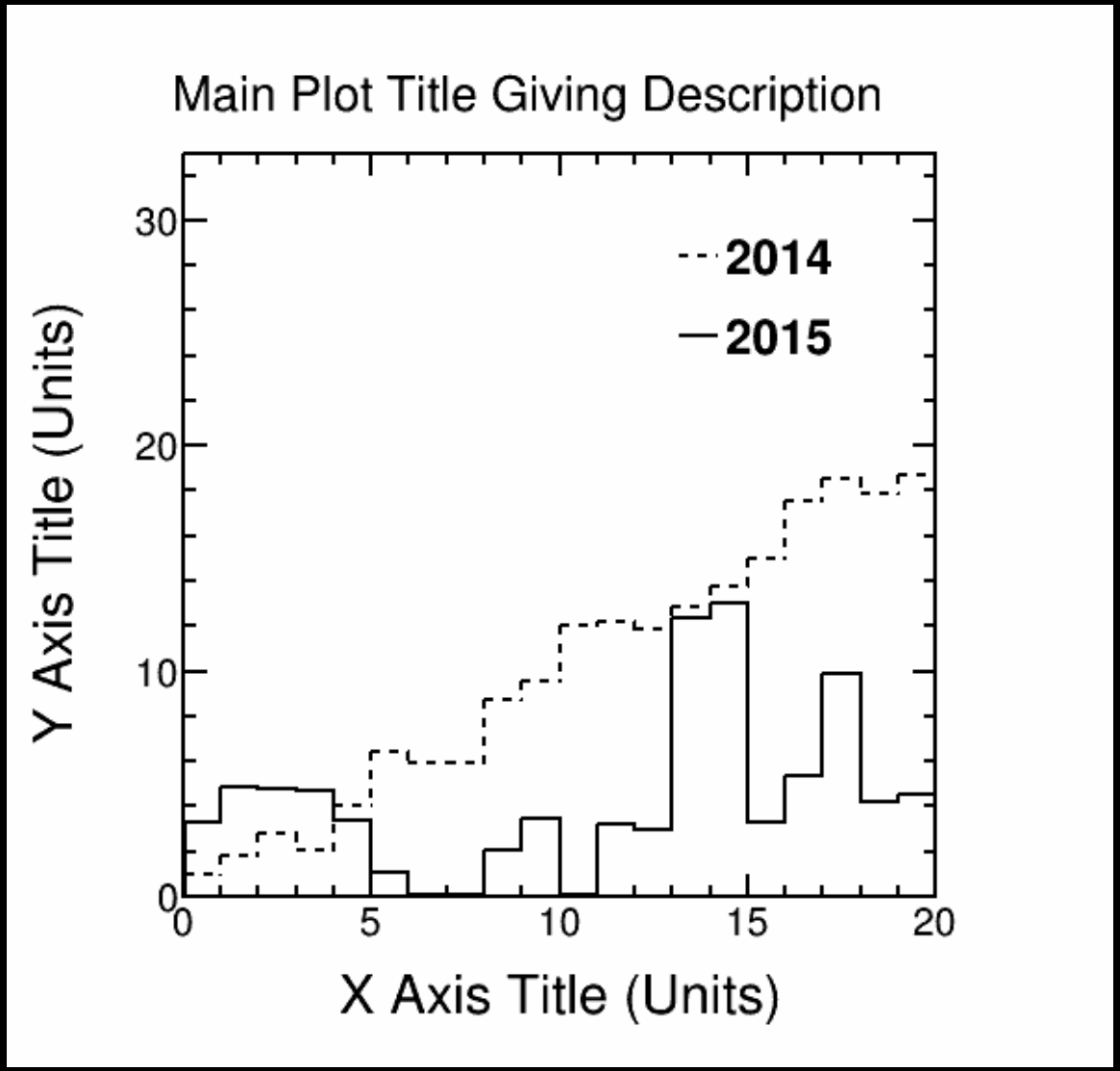
Types of Questions

- begin given a question and asked to describe/sketch the right plot to answer it
- write R psuedo-code to solve a problem similar to those met in the homework
- asked to describe relationships
- asked to discuss/design experiments
- …

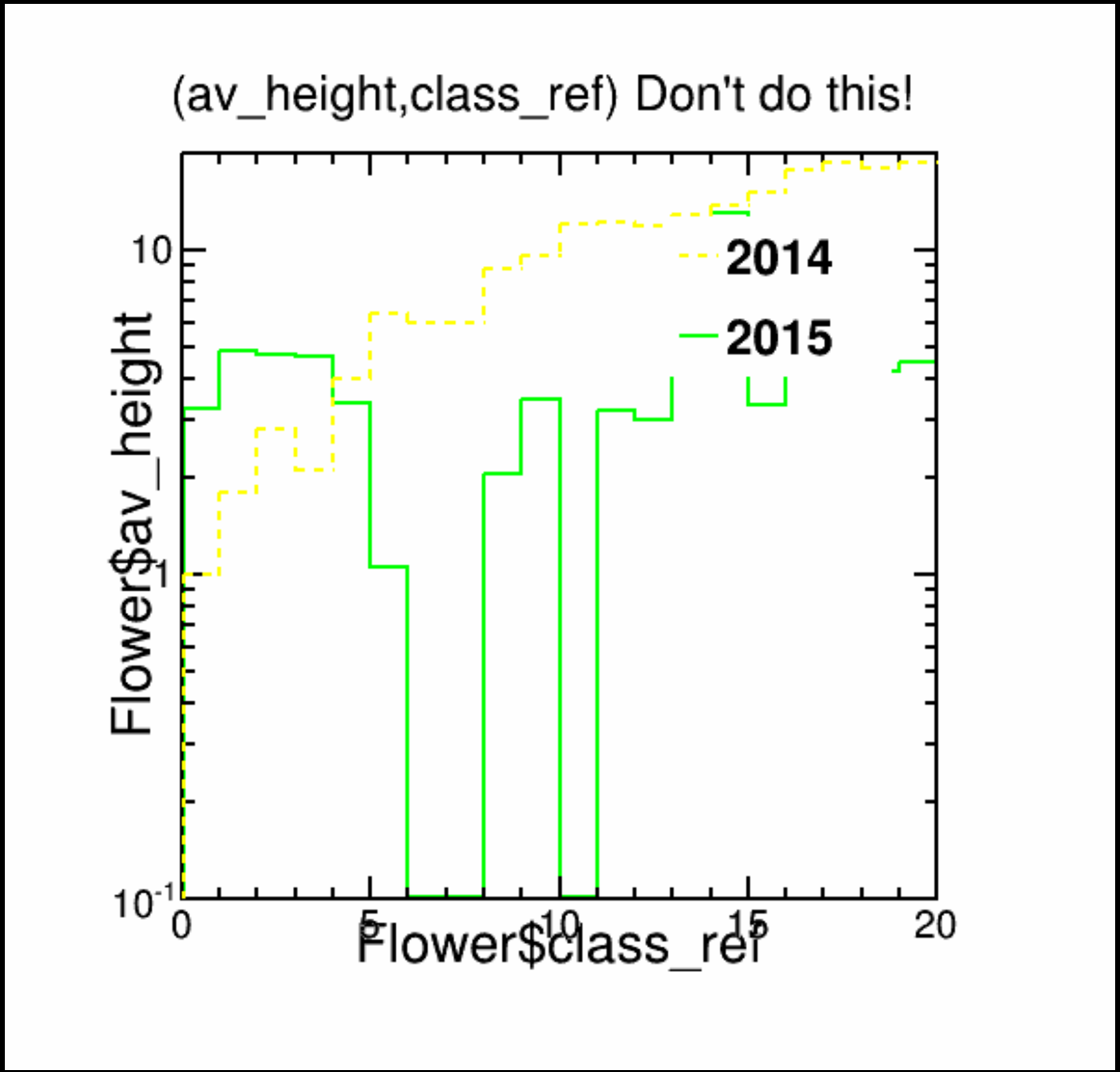
A Good Plot

- Title
- Axis titles
- Numbers and units on all axes
- Legend labelling all lines if more than one
- Clear what is plotted
- Legible
- Colours visible on projectors and in print

Plot Outline



Bad Plot...


(av_height,class_ref) Don't do this!
2014
2015
Flower$av_height
Flower$class_ref
10
1
10^-1
0
5
10
15
20

What does it mean to make the right plot?

- Ask a good question
- Answer the question in one plot

Correlation

Strong/Weak?

Direction?

Linear/Non-linear…?

How to discover correlations?

Make some plots!! Of course…

- Typically make scatterplots of each pair of variables
- Can you see a relationship? If not then not strong.
- Describe relationships using functions (not only linear!)

Function Examples

- linear, quadratic, cubic…polynomial
- exponential, logarithmic
- Gaussian
- etc.

Why we need to be quantative…

Later on we are going to try to use some variables to predict others, this requ ires fitting a sensible function to the available data.

These problems come in two main categories:

1. You have a theoretical model for how the variables should be related
2. You have no theoretical model for the relationship and you have to guess somehting from the data
3. Some combination of the two due to *e. g.* unexpected noise.

The second case: guessing functions

Often there is not a single right answer (theme of this course…)

Which function is good enough?

- Needs to describe the major features of the data
- Should be minimal - as simple as will work
- May well not be unique, you can try fitting different functional forms and see which one works best
- We will start the actual fitting next week

What is a feature?

Things to look for and check match:

- behaviour as x $\rightarrow \pm\infty$
- turning points (gradient = 0)
- crossing points with the axes

Types of Causation

If A and B are correlated then there are different possibilities of causation:

- A causes B
- B causes A
- C causes A and B (lurking factor)
- A causes C which causes B (or vice versa)
- A causes B and B causes A (cyclic or bidirectional)
- there is no connection between A and B only coincidence

Interpolation/Prediction

Interpolation - finding values between two known points

Prediction - estimating values outside the range of the data

Fitting vs Interpolation

Interpolation - when you have no or very small errors, for example in your model calculations

Fitting - to deal with errors in your data

Regression Fitting

We want to get a function that describes our data well, but we know there are uncertainties in the data causing some scatter in the data points

Under and Over Fitting

If you define a suitably complex function, you can get it to pass through all of your data points (like with the splines). However, this does not mean that the features in you function really exist, rather they are probably caused by statistical noise - **over fitting**.

If you try to fit a straight line to a non-linear functional relationship then you will not be able to well describe the behaviour of the data - **under fitting**.

Checking for under/over fitting

1. Look at the data and fit and use your brain
2. Ask if the function you have used is the simplest one that could describe the data?
3. Ask if the fit describes the data well?
4. Test the fit on a subset of the data (training set)

Applying Statistics to Distributions

Experimental Design

Exam Advice

Stay calm and think.

Ask yourself: does my answer make sense?

Exam Advice

Also revise statistics, writing R psuedo-code!

Good Luck!

Project Presentations

Backup Slides