4.7 Exercise: Techniques for scatterplots

This exercise will enable you to become more proficient in creating scatterplots with iNZight. You will learn how to apply the most suitable trend line and use techniques to overcome perceptual problems.

The skills addressed are:

1. Create a scatterplot of two numeric variables and apply a suitable trend line.
2. Use techniques such as jittering, transparency and running quartiles to deal with overprinting.

[*iNZight Lite version linked here*]

INSTRUCTIONS

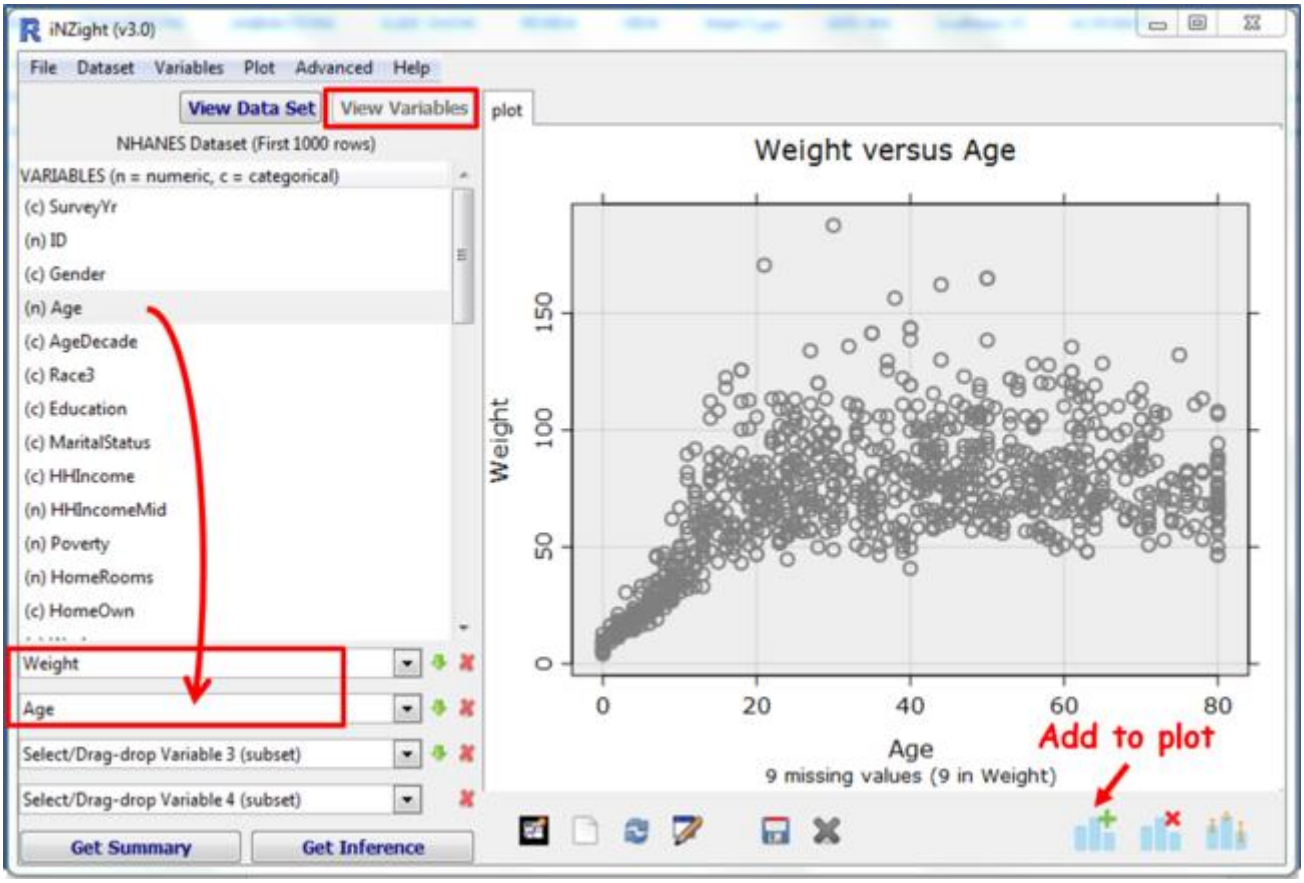
Follow the instructions below to generate the graphs. Or you may prefer to print these instructions. If you have a problem doing the exercise, scroll down to **Common questions**.

Load the **nhanes_1000** dataset into iNZight using **File > Example data ….** You will find the data set in **Module (package)** *FutureLearn*.

Choosing a suitable trend curve or smoother

We are going to explore the relationship between variables **Age** and **Weight** of people in the **nhanes_1000** dataset.

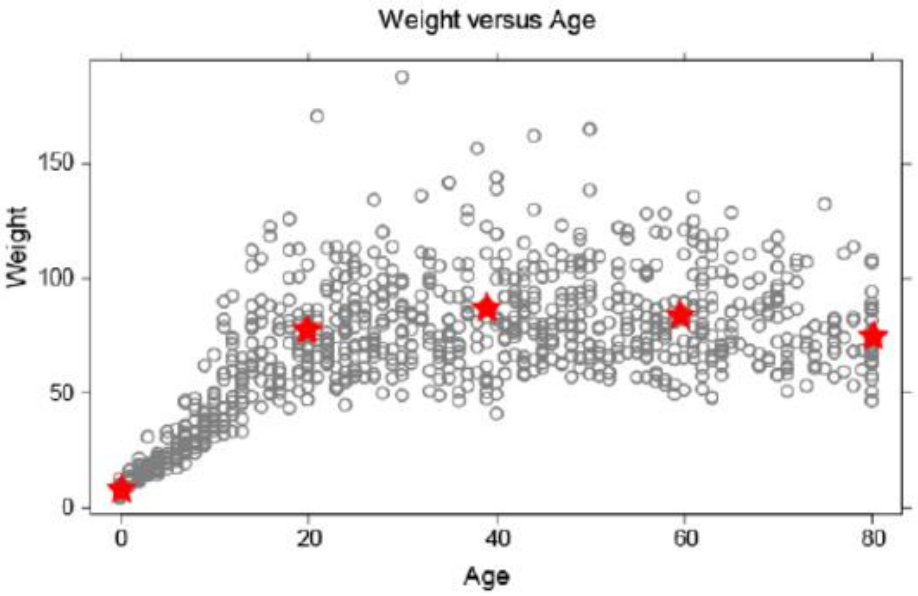
Construct a scatterplot of **Weight**, the outcome variable, and **Age**, the predictor variable.



Take a little time to look at the graph and think about it in terms of **centre**, **spread**, **shape** and **oddities**.

If you think that there is a relationship between **Age** and **Weight**, how would you describe it to someone?

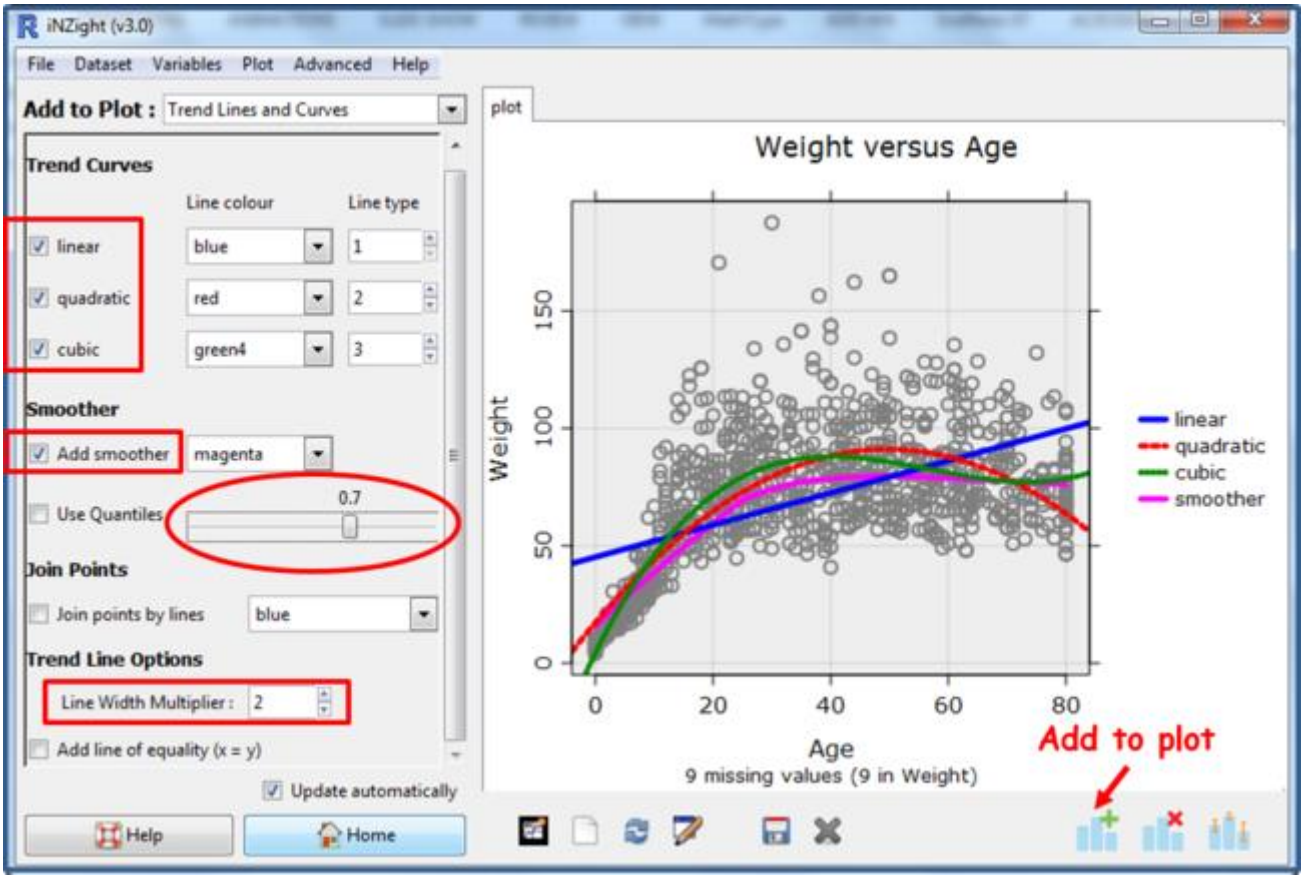
Where do you think you would sketch a trend line? Would it be curved or straight? I have placed stars in the centre of the data for each 20 years.



In **Add to Plot** select **Trend Lines and Curves**. By clicking the relevant check boxes, add the four available trend curves. You can adjust the thickness of the lines using the **Line Width Multiplier**. Now adjust the **slider** which determines how smooth or wiggly the smoother is.

Which trend curve do you think fits the data the best?

Post a comment stating your choice of trend curve and explain why.



The points tend to overwhelm the added trends in this picture because the grey is so dark. It would be useful to make the points smaller and lighter.

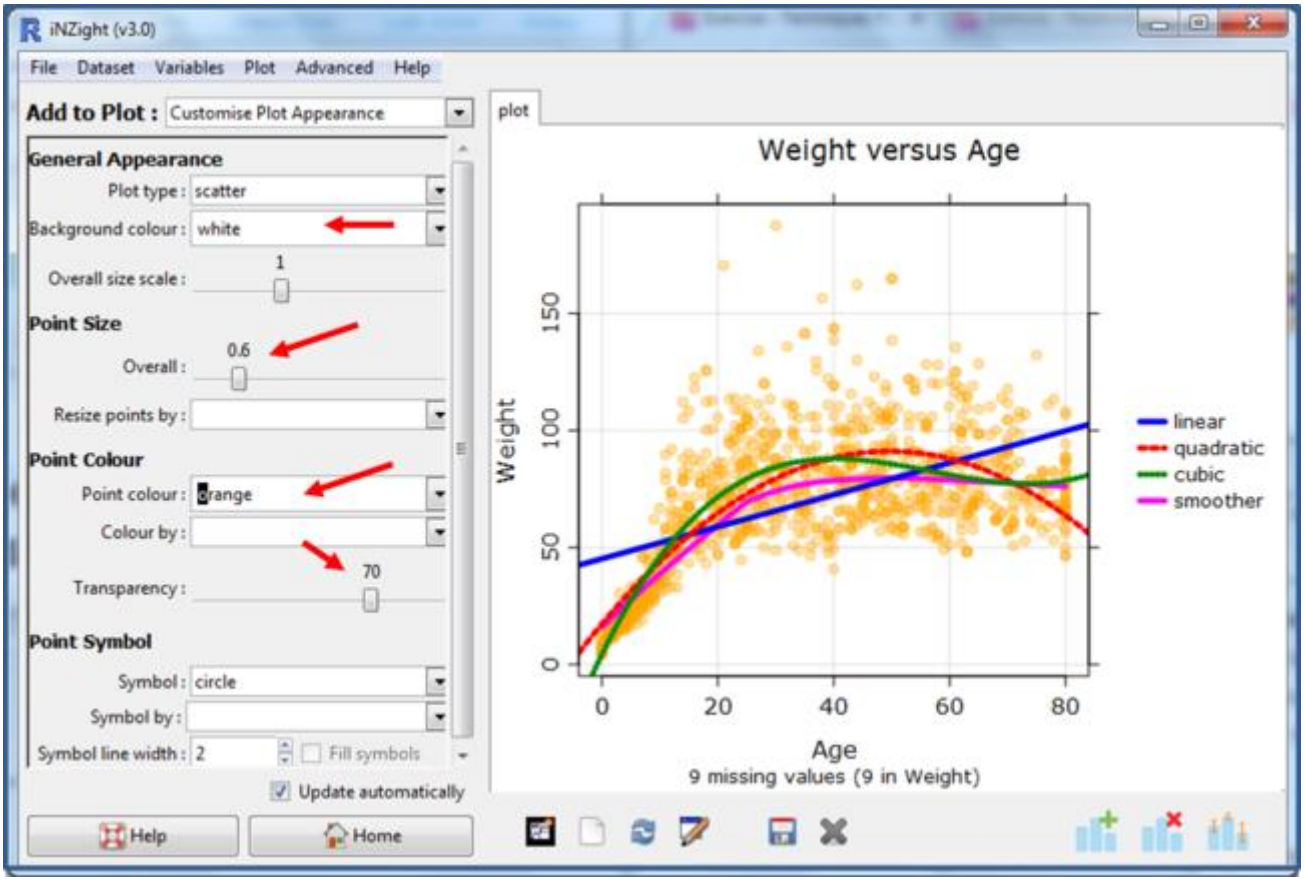
Techniques to deal with overprinting

Whenever we graph a large dataset some values will be printed over each other. We may place too much emphasis on the small numbers of values scattered around the outside edges of the plot. As discussed in the video, we can use **transparency**, **running quartiles**, and **jittering** to get a clearer picture of the density of the data.

Transparency

To see where there is a lot of overprinting, we can make the points transparent. To make the points transparent:

- Still in **Add to Plot** select **Customise Plot Appearance**
- Move the slider next to **Transparency** .



What do the darker spots on the above graph represent?

PRACTICE: (~5min)
In the plot above we have reset the **Background colour** to *White*, reduced the **Point size** to 0.6, made the **Point colour** *orange* (by simply typing that colour name into the box because it is not in the dropdown list), and slid **Transparency** to 70(%).

Play with settings like this until you get something you think works well. You can also go back to **Trend Lines and Curves** and adjust trend-line types and colours.

Running quartiles

Add a smoother to your graph that goes through the median weight for a given age.

- In **Add to Plot** select **Trend Lines and Curves**
- Deselect Linear, Quadratic and Cubic so only **Add smoother** is selected. [Note: We changed the smoother colour to black so it is easier to see.]
- Now click **Use Quantiles**.
  This gives a running Median and 1st and 3rd quartiles of the data.

Note: If you have a larger dataset you will also get 10th and 90th percentile lines.

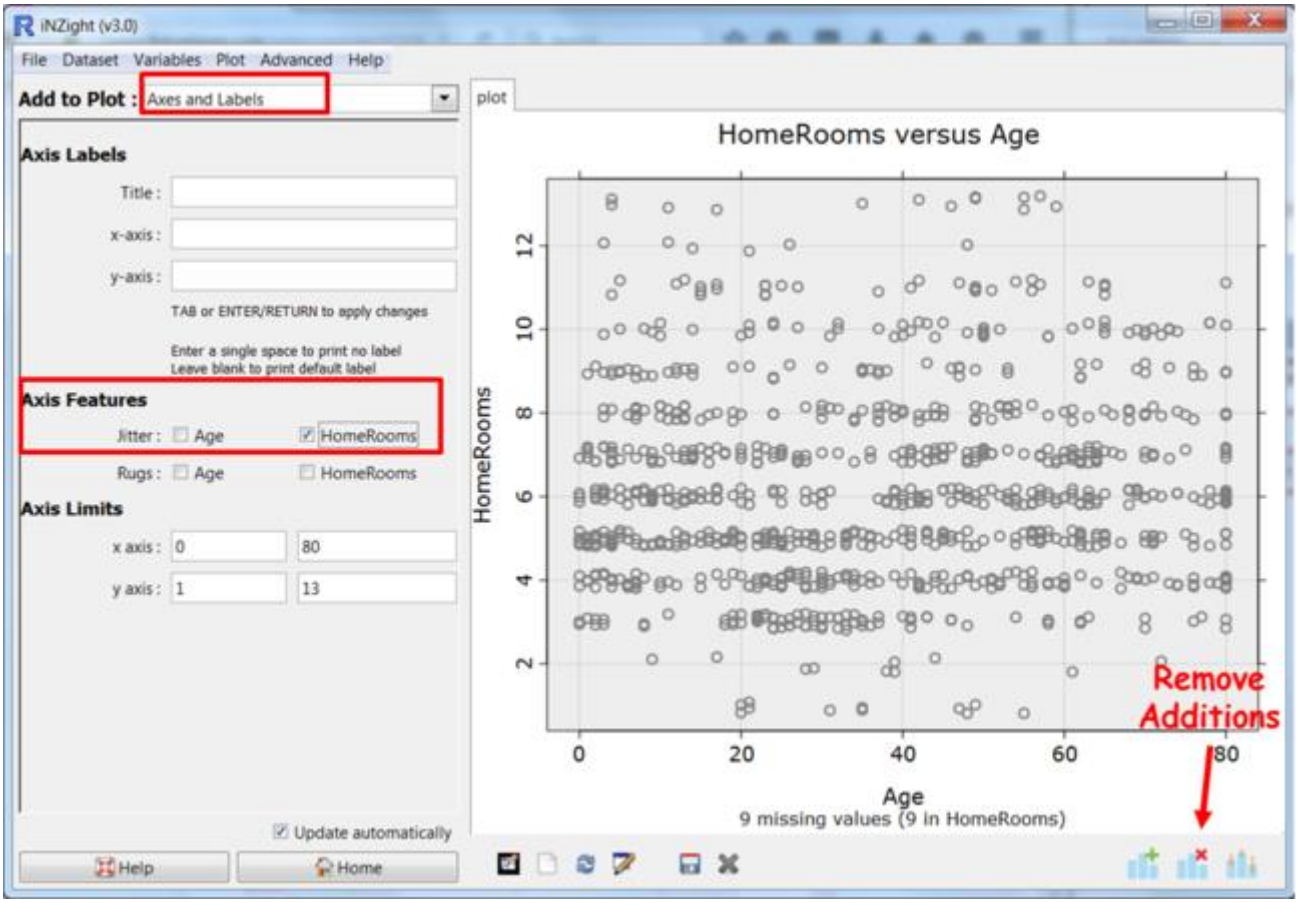
What do the dotted lines mean? What can we say about the weights and ages of the people in the **nhanes_1000** dataset? **Post a comment if you see something interesting.**

Jittering

Jittering is a useful technique when we have a lot of overprinting, especially when we have discrete numeric variables such as the number of rooms at home.

Using the **nhanes_1000** data, create a new plot of **HomeRooms** versus **Age**. This plot will appear with a lot of points overlapping in straight horizontal lines. [You may have to use **Remove Additions** and **Remove all additions** to get rid of all the colouring etc. we have been using.]

- In **Add to Plot** select **Axes and Labels**. Under **Axis Features**, see **Jitter** .
- Ask for **Jitter** on **HomeRooms**.



With jitter added you should see the points that were previously overprinted.

Optional: *Try this new feature* (*interactive web graphics*)

Save your graph as **File Type:** *Interactive HTML* (you will have to supply a name for the file). The file will open up in your default browser. If that is a modern browser like Chrome, Firefox or Safari (but not Internet Explorer) this will then give you an interactive version of the graph that lets you query it in various ways like hovering over the points, or a trend line, or clicking them, or selecting more than one using the Ctrl or Shift keys, or by dragging.

The save process can be slow if there are a lot of dots to be drawn.

The save window allows other variables to be exported along with the plot. This is particularly useful for hover-over if you have a variable that gives the names of the people or objects.

You can give such files to others. They do not need to be connected to iNZight to work.

Common questions

Which variable do I jitter?
If you have a variable that has a lot of very common discrete values (e.g. **HomeRooms**) with gaps between them, (horizontal or vertical white space) you should jitter that variable. Look at the axes and see whether it is the variable on the x (horizontal) axis or the y (vertical) axis.

Is a smoother a trend curve?
Yes, a smoother is a way of estimating a trend curve.

How wiggly should it be?
Be guided by your eye. Think in terms of vertical slices and trying to keep the trend in the centre (vertically) of the data points.