4.7 Exercise: Techniques for scatterplots *(iNZight Lite version)*

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.

INSTRUCTIONS

Import the **NHANES-1000** dataset into iNZight Lite:

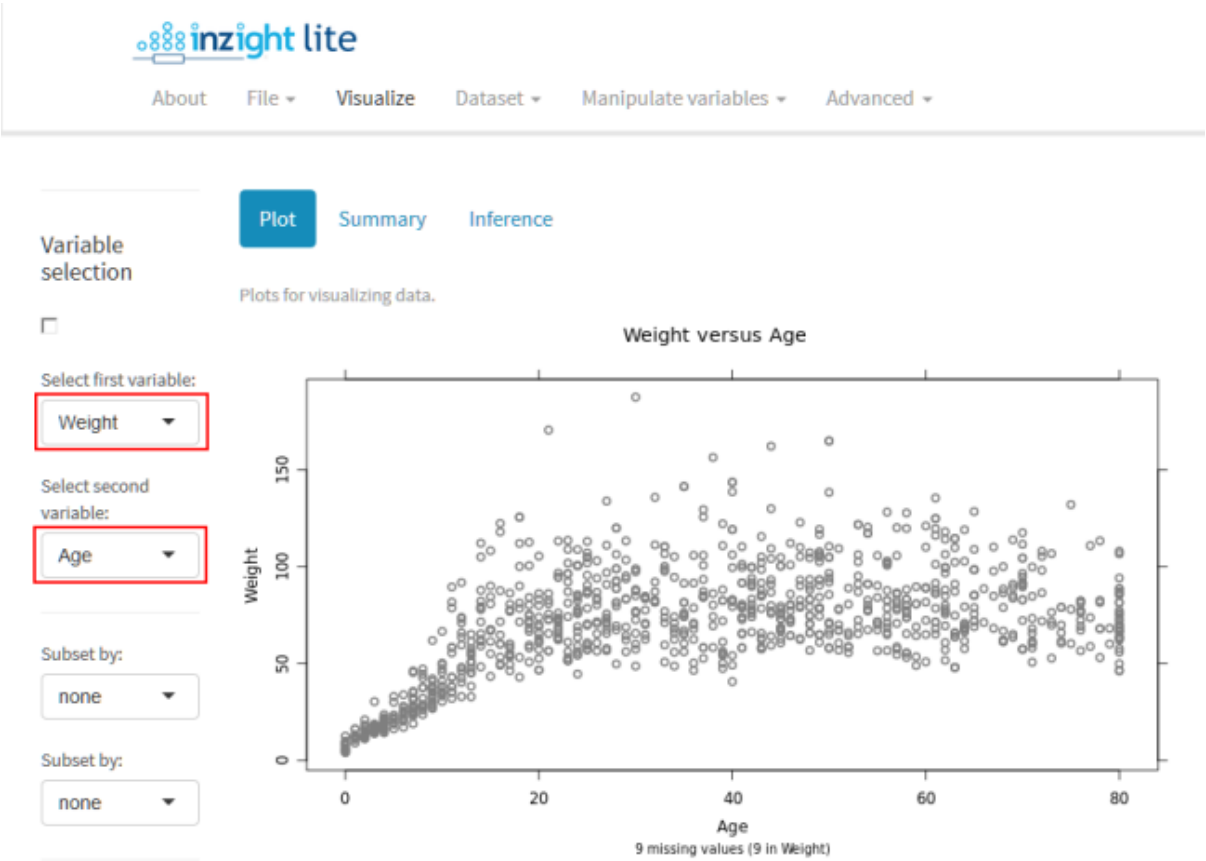
- Select **File** > **Dataset Examples**
- Select Data set category: **Future-Learn**
- Select **NHANES-1000**
- Click on **Select Set.**

If you have any problems during this exercise, see **Common questions** on page 9.

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.

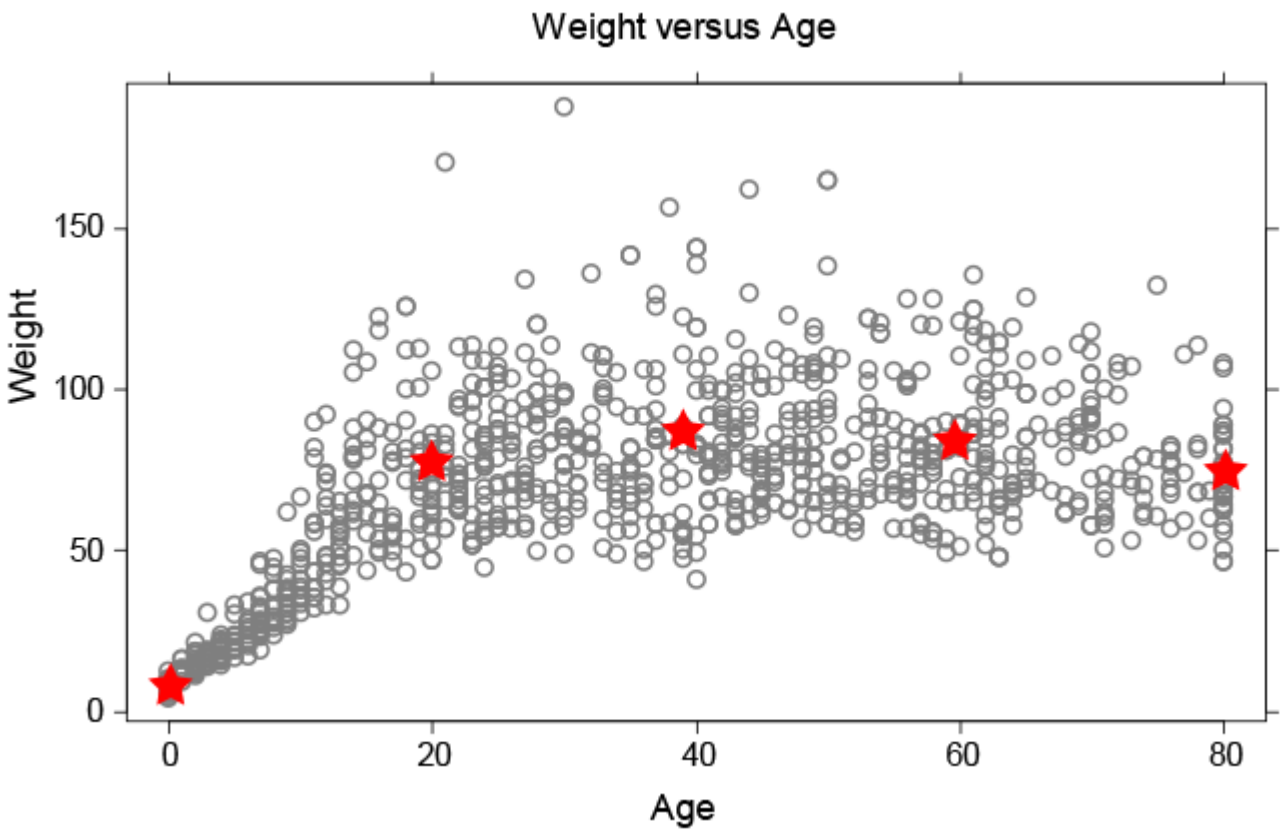
Construct a scatterplot of **Weight**, the outcome variable, and **Age**, the predictor variable.



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.



Select Add to Plot and choose **Trend Lines and Curves**. Try adding the four different trend curves. 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.

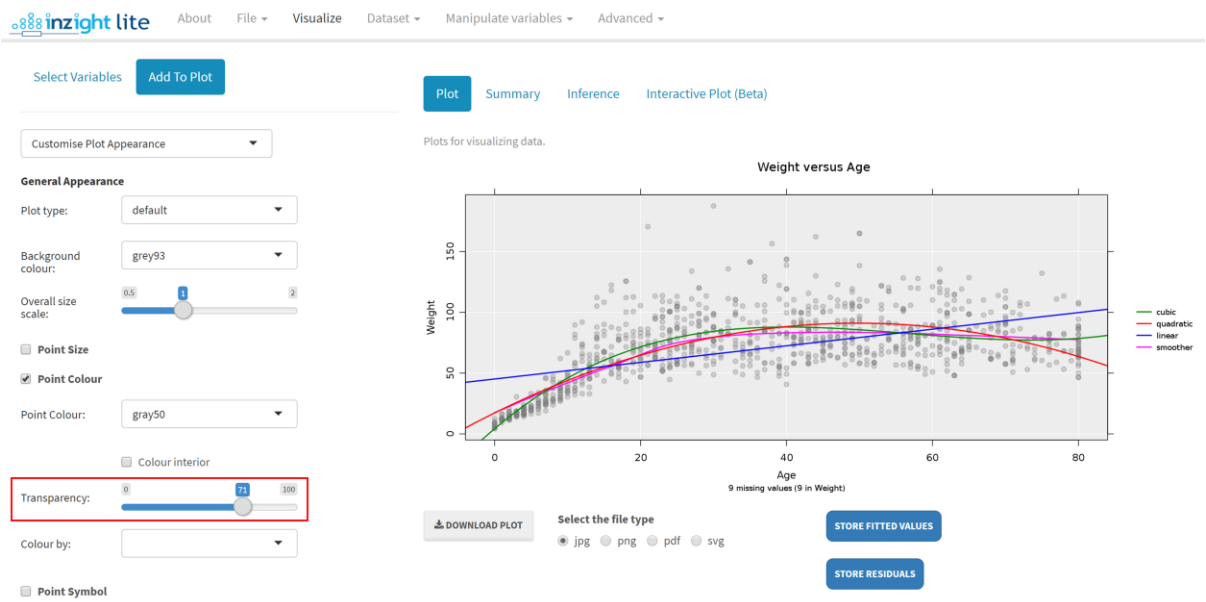
Techniques to deal with overprinting

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:

- Select **Customise plot appearance**
- Click **Point Colour** to reveal the options
- Move the slider next to **Transparency**



What do the darker spots on the graph represent?

PRACTICE: (~5min)

Play with the levels of **transparency**, **colour** and **size** of your symbols until you are familiar with the features of **Change Plot Appearance.**

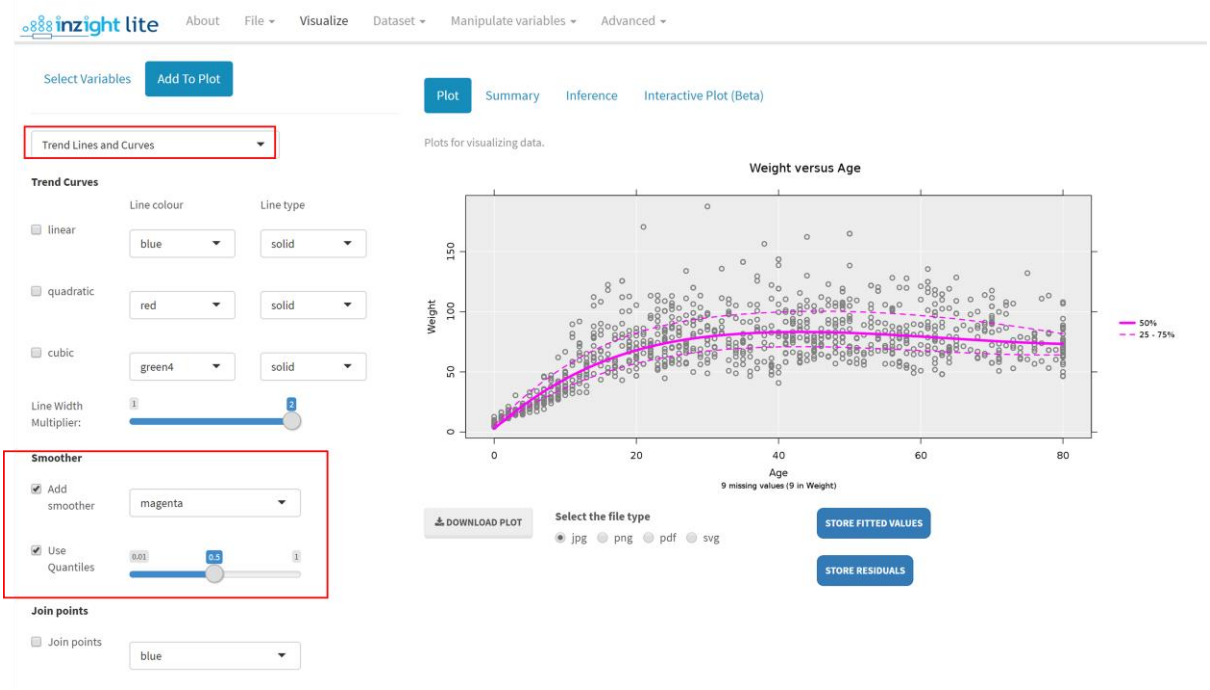
Running quartiles

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

- Select T**rend Lives and Curves**
- Tick **Add smoother**

Now add quantiles which run through the 1st and 3rd quartiles of the data

- Tick **Use Quantiles**



Notes: If you have a larger dataset you will also get 10th and 90th percentile lines. If your quantile lines are too faint use the **Line Width Multiplier** above the Smoother option.

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

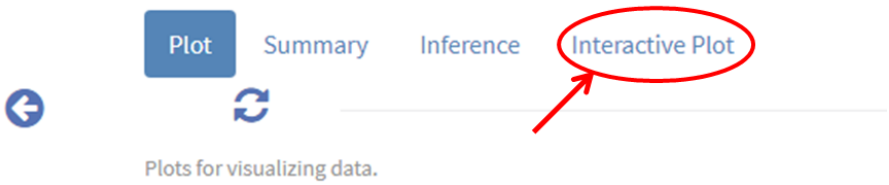
Jittering is a useful technique when we have a lot of overprinting, especially when we have discrete numeric variables such as the number of rooms at home.

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.

- On **Add to Plot**, select **Axes and Labels**
- Under **Axis Features** click the **Jitter HomeRooms (y-variable)**

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

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



Click on the **Interactive Plot** tab. This will give you an interactive version of your 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. You have to Click **Produce Plot** for it to appear. This is a protection as the conversion process can be slow if there are many points to be drawn.

A click button (**Select additional variables**) ***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 plotted.

You can download these plots as Interactive HTML files which you can give to others. They do not need to be connected to iNZight Lite 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, we refer to a smoother as a trend curve.

When I used a smoother with Use Quantiles clicked I lost the running quartiles when I changed the line colour?
A small bug. Uncheck **Use Quantiles** and the click it again.

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.