

3.5 Exercise: Relationships between categorical variables

This exercise will enable you to construct graphs of two categorical variables as discussed in the previous video. The skills addressed are:

1. Creating a plot of two categorical variables (when the predictor variable has only 2 groups).
2. Making a side by side bar chart of two categorical variables.
3. Filtering out unwanted groups within a category.
4. Graphing a predictor variable with more than two groups.

[iNZight Lite version [linked here](#)]

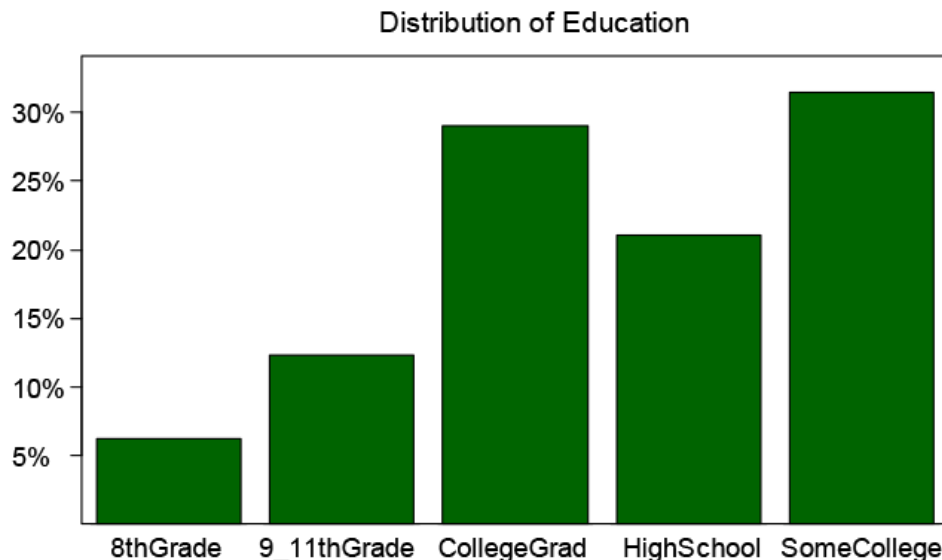
INSTRUCTIONS

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**.

To begin this exercise load the **nhanes2009-2012** dataset into iNZight using **File > Example data** You will find the data set in **Module (package) FutureLearn**.

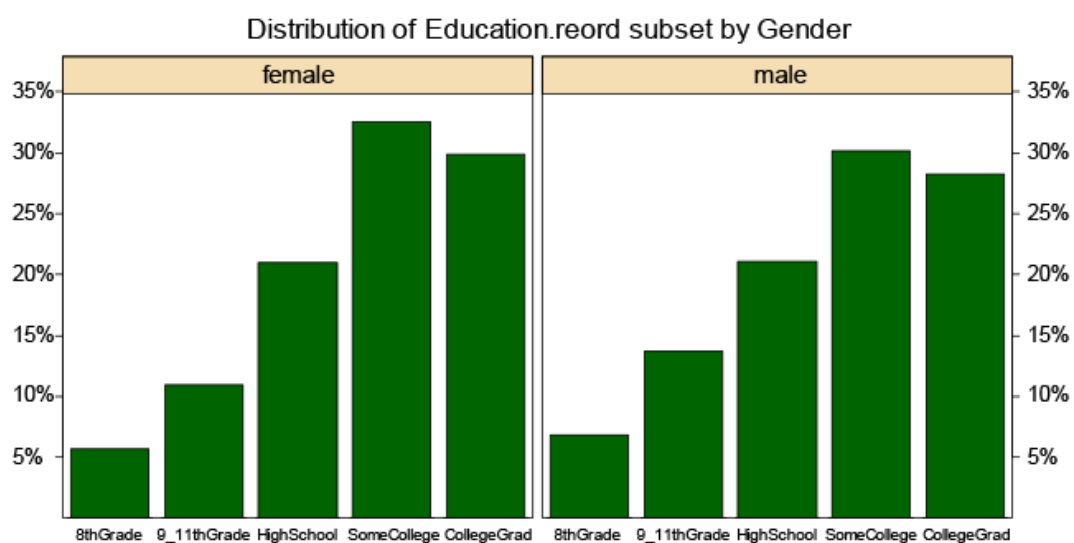
Plot two categorical variables

We are interested in how gender affects educational attainment so our outcome variable, **Education**, should go in the **Variable 1** slot.

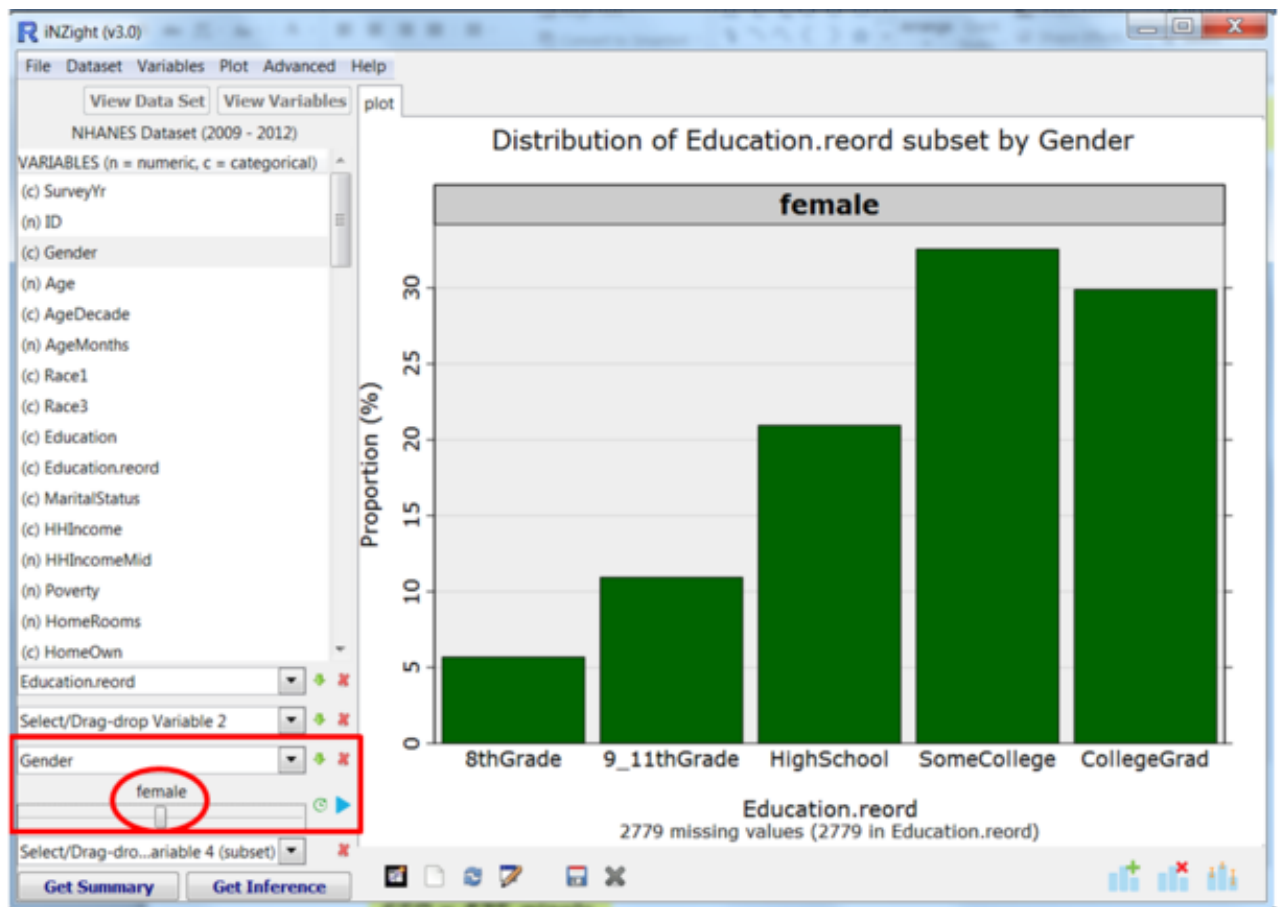


We see the plot, but the bars are in alphabetical order, rather than level of educational attainment. We need to manipulate the variable so that we see the categories of education in a more useful order. Exercise 2.5 showed the use of this technique to create a new variable called *Education.reord*. You will need to do that again.

Drag **Education.reord** to the **Variable 1** slot and our predictor of interest, **Gender**, to **Variable 3 (subset)**. You get separate plots for the female and male subsets (subgroups).



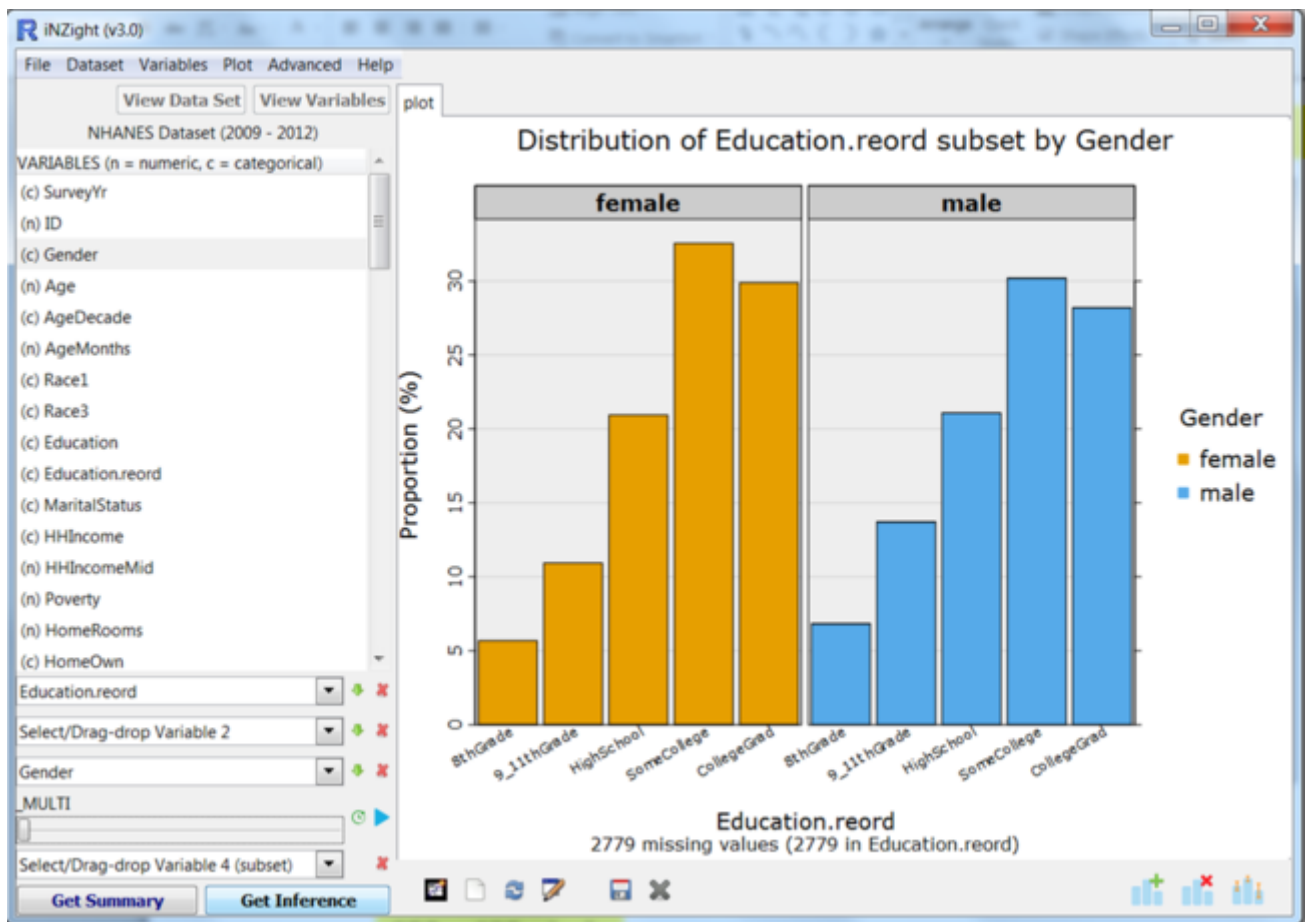
Drag the slider to the right. You will see individual graphs of Educational Attainment for females and males.



PRACTICE (~5 min)

Are there any more interesting categorical variables in the data set that may be associated with education? Try dragging other variables into the **Variable 3 (subset)** slot.

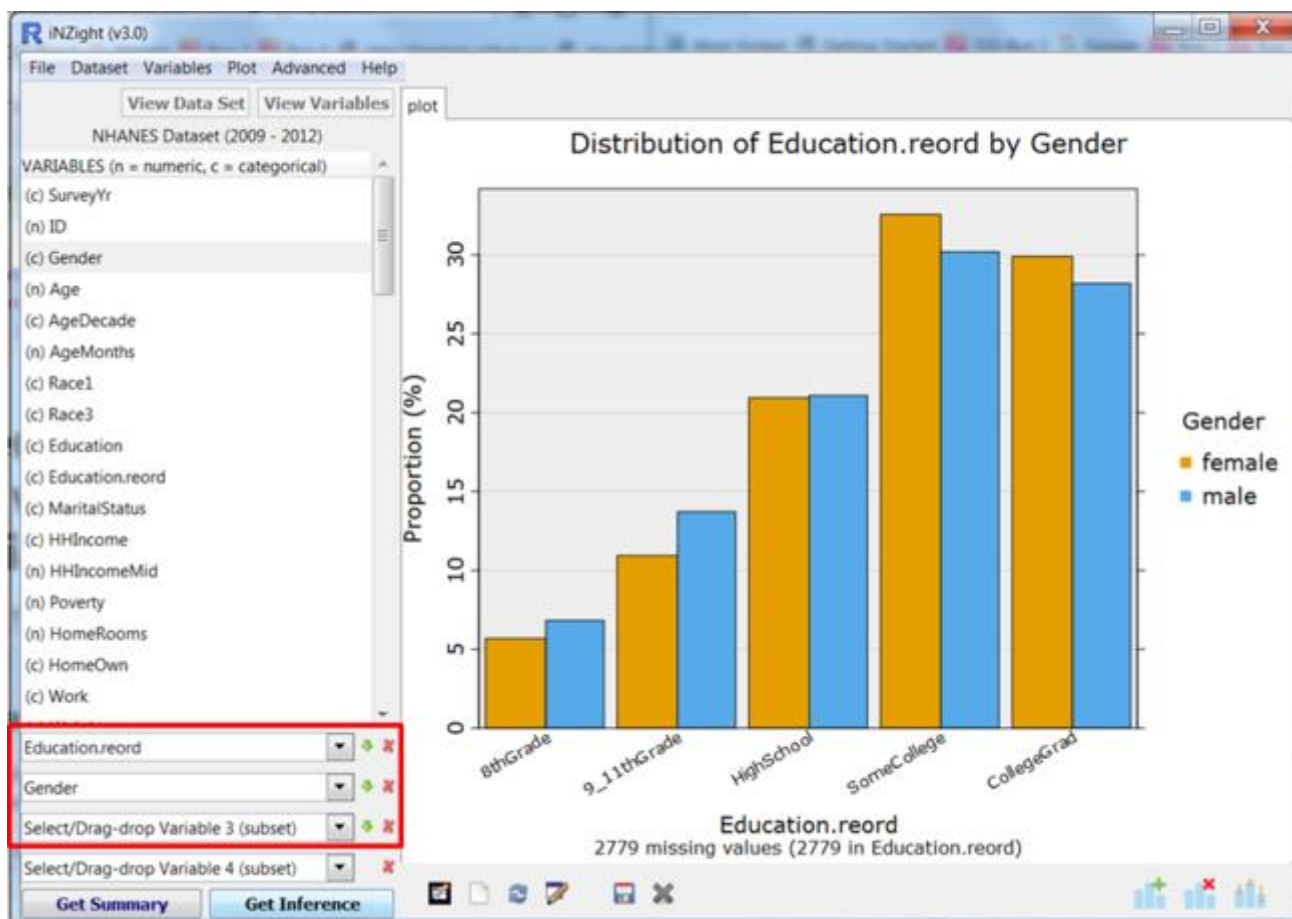
When you have finished, put **Gender** back in **slot 3**. Also **colour by Gender** using the Colourblind friendly palette (Find "Colour by" under "Add to Plot").



Side by side bar charts of two categorical variables

As we heard on the video, if we want to have a closer look at the differences between females and males for each level of education, we can create a side by side bar chart.

Clear **Gender** from the **Variable 3 (subset)** slot. Then drag **Gender** into the **Variable 2** slot and keep **Education.reord** in the **Variable 1** slot. [Speed Tip: Use the little green arrow next to Slot 2. It interchanges the contents of Slots 2 and 3.]



By clicking the little green down-arrow alongside Slot 2, shuffle between the side-by-side-bar-chart and separate-bar-charts presentations of these data. *[This interchanges the contents of Slots 2 and 3.]*

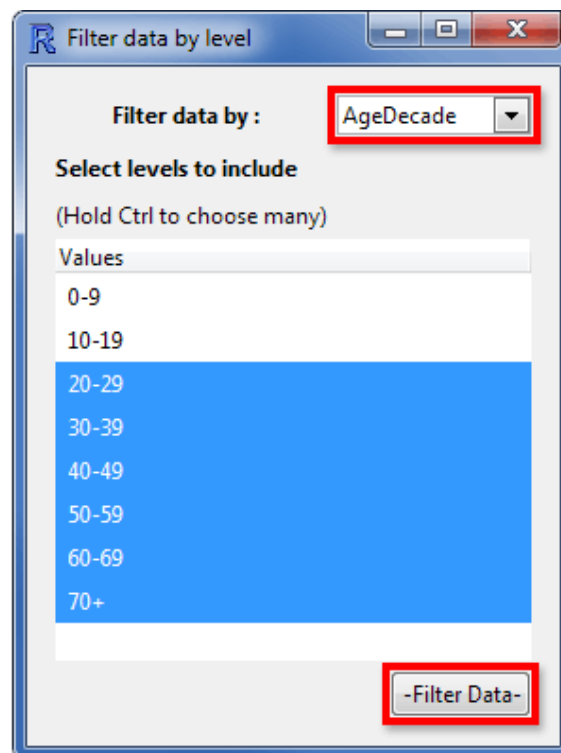
Filtering out unwanted groups within a category

We hope it has been bothering you all this time that there are a large number of missing values for these plots. Why? If you look at the documentation for the data sets, Education is not recorded for those aged under 20. We will need to filter out the under 20s when we look at Education (and should have done so from the outset).

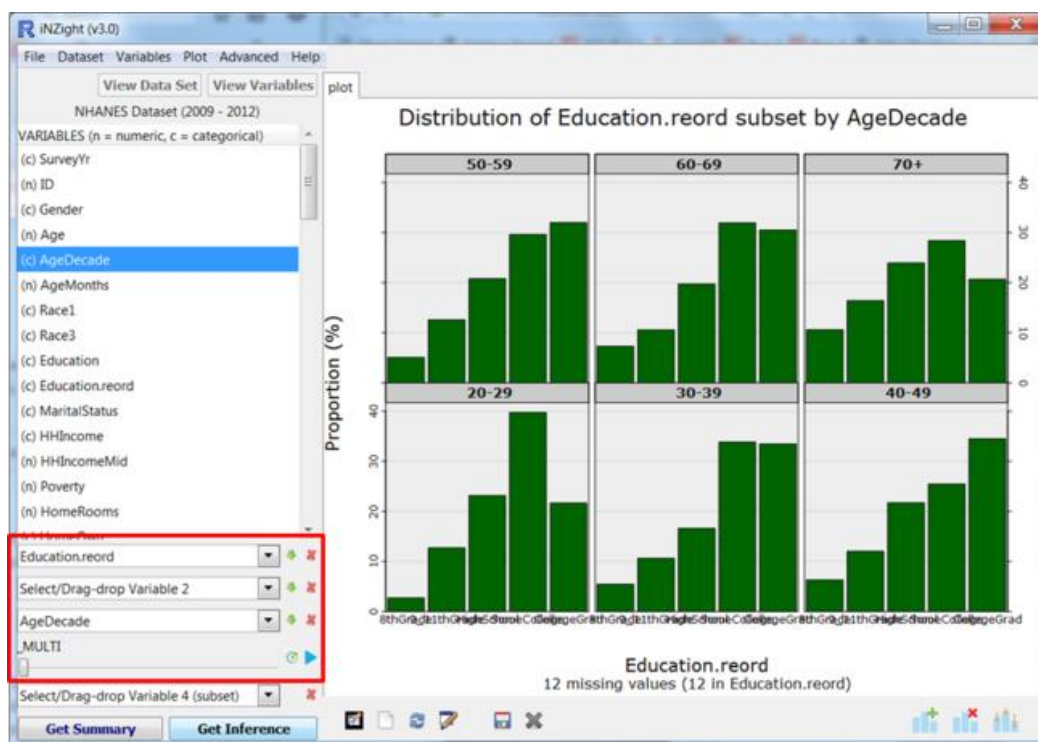
From the top menu, go **Dataset > Filter Dataset**

- Select **Levels of a categorical variable** and then click **Proceed**.
- Select the variable you wish to filter, e.g. **AgeDecade**.
- Hold the **CTRL** key on your keyboard and select **ALL** of the categories you wish to **include**, e.g. all groups representing ages 20 and older. (Note: The **Shift** key lets you select a whole set of groups with no gaps between them.)

- Click **Filter Data**.



The dataset will now include only the selected age groups and you can recreate the graph for **Education.reord** subset by **AgeDecade**.



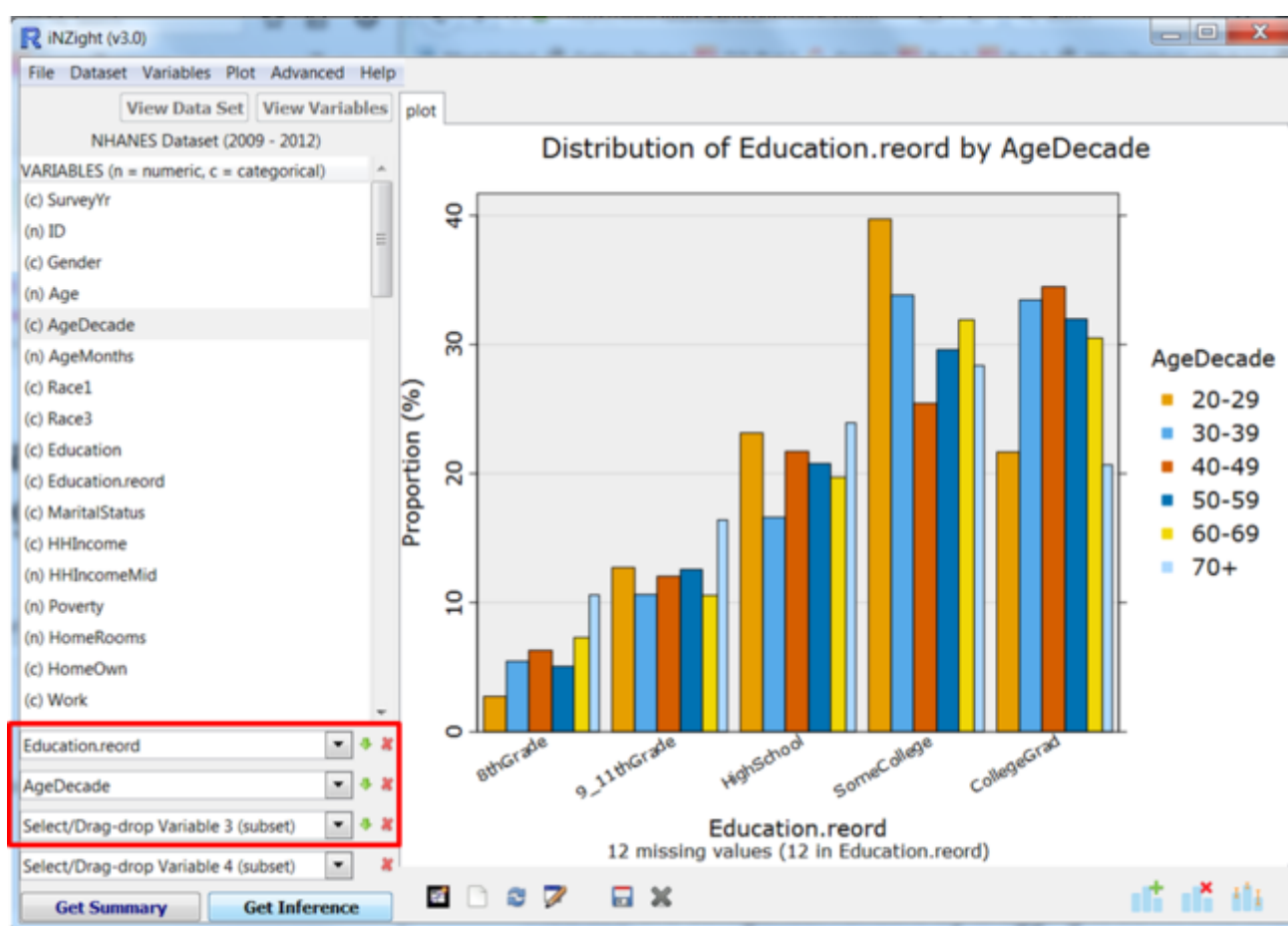
You will see that the under 0-9 and 10-19 groups do not appear. We have reduced the number of missing values from 2779 to 12.

PRACTICE (~5 min)

Play through the graphs using the **play** button. Use the slider to look at individual graphs.

Compare the shapes of your plots. Post your findings in the comments area.

Now create side by side graphs for each level of **Education**. Clear **AgeDecade** from the **Variable 3 (subset)** slot. Then drag **AgeDecade** into the **Variable 2** slot (or just click the green arrow beside Slot 2).



Are there any more interesting predictor variables in the data set?

PRACTICE (~5 min)

Add in a 3rd categorical variable by dragging it down into the **Variable 3 (subset)** slot. Practise moving the slider between graphs.


Optional

Try this new feature (Interactive plots)

[Warning: At present this feature only works for single graphs and not when you have multiple graphs on the same page.]

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 bars or clicking them. Explore!

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

Alternatively when this icon  appears and is blue, simply clicking it will fire up an interactive version of the current plot.

Other ways of representing relationships between 2 categorical variables (iNZight versions from 3.4.6)

There are several ways of plotting relationships between 2 categorical variables. Go to **Add to Plot** and look at what is delivered by the various options under **Plot type**. Can you see relationships between the ways the various types of graph represent the information? Play with some of the controls for each plot type.

Common questions

I no longer want my dataset filtered, how do I get all of my values back?

Dataset > Restore Dataset gets you back the data set as originally imported.

My graph is not showing up properly in my plot window.

Click **Redraw plot** 3rd icon from left at the bottom of the plot window (or **Plot > Redraw plot**).

In the panel of graphs the Age groups are ordered from bottom-left to top-right. Why is that?

They are behaving like numbers plotted on a scatter plot – they get larger going up the page and towards the right.