3.5 Exercise: Relationships between categorical variables (*iNZight Lite version*)

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.

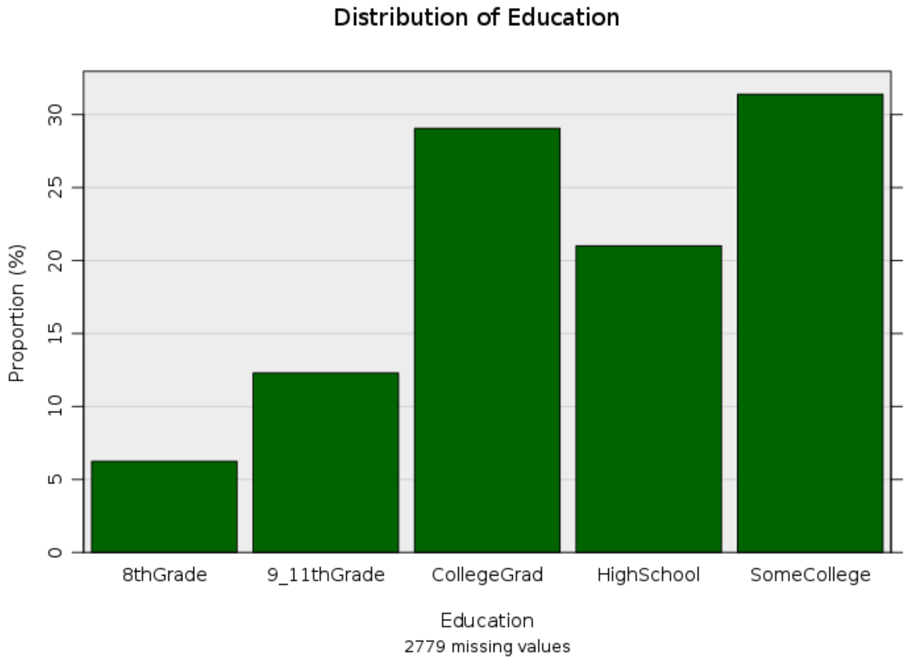
To begin this exercise, load the dataset example **NHANES 2009-2012**.

Import the **NHANES 2009-2012** dataset into iNZight Lite:

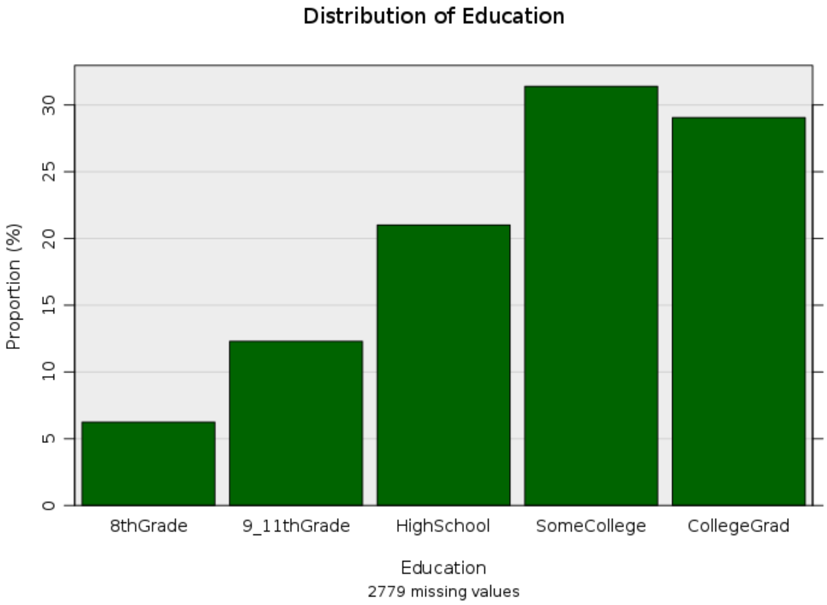
- Select **File** > **Dataset Examples**
- Select Data set category: **Future-Learn**
- Select **NHANES 2009-2012**
- Click on **Select Set.**

Plot two categorical variables

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



We see the plot, but the bars are in alphabetical order and that makes it harder to read the graph. Now **put the categories of Education** into the **natural order**. (Exercise 2.5 showed how to do this).



Leave **Education** in the **first variable** slot and select our predictor of interest, **Gender**, in the first **subset by** slot.

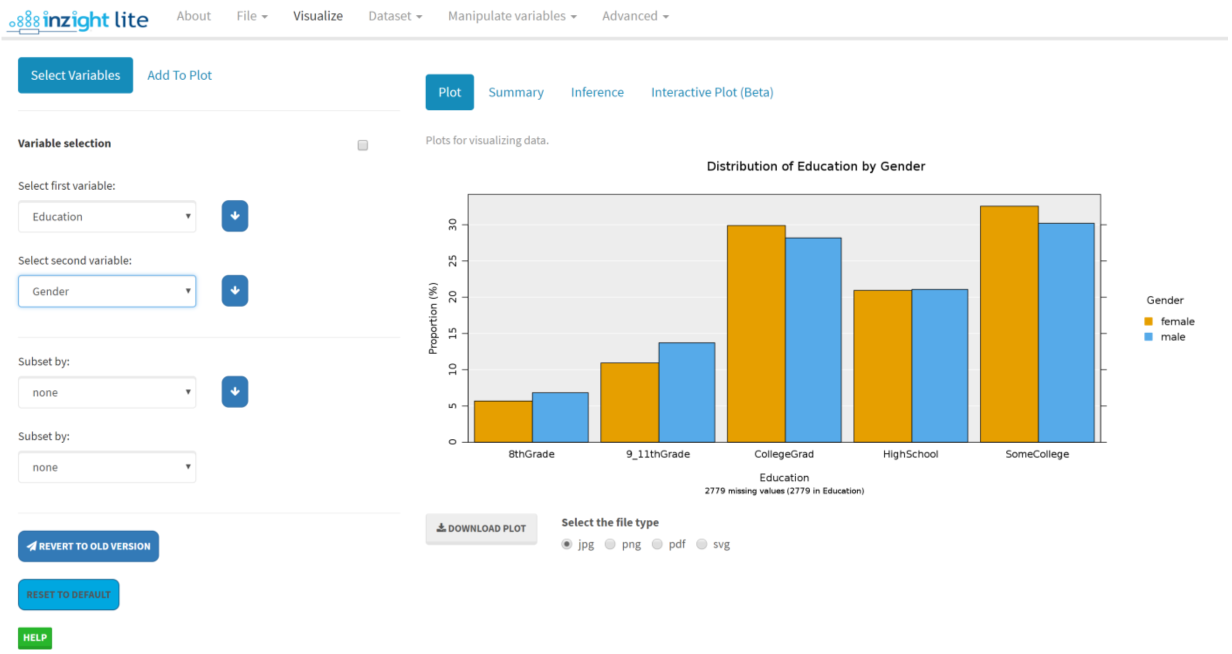
Drag the slider to the right. You will see individual graphs of educational attainment for females and males.

PRACTICE (~5 min)

Are there any more interesting predictor variables in the data set? Post your findings on the discussion.

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 **subset by** slot (Select none). Then select **Gender** in the **second variabl e** slot and keep **Education** in the **first variable** slot.



Variable selection

Select first variable: Education

Select second variable: Gender

Subset by: none

Subset by: none

Select the file type

● jpg ○ png ○ pdf ○ svg

Filtering out unwanted groups within a category

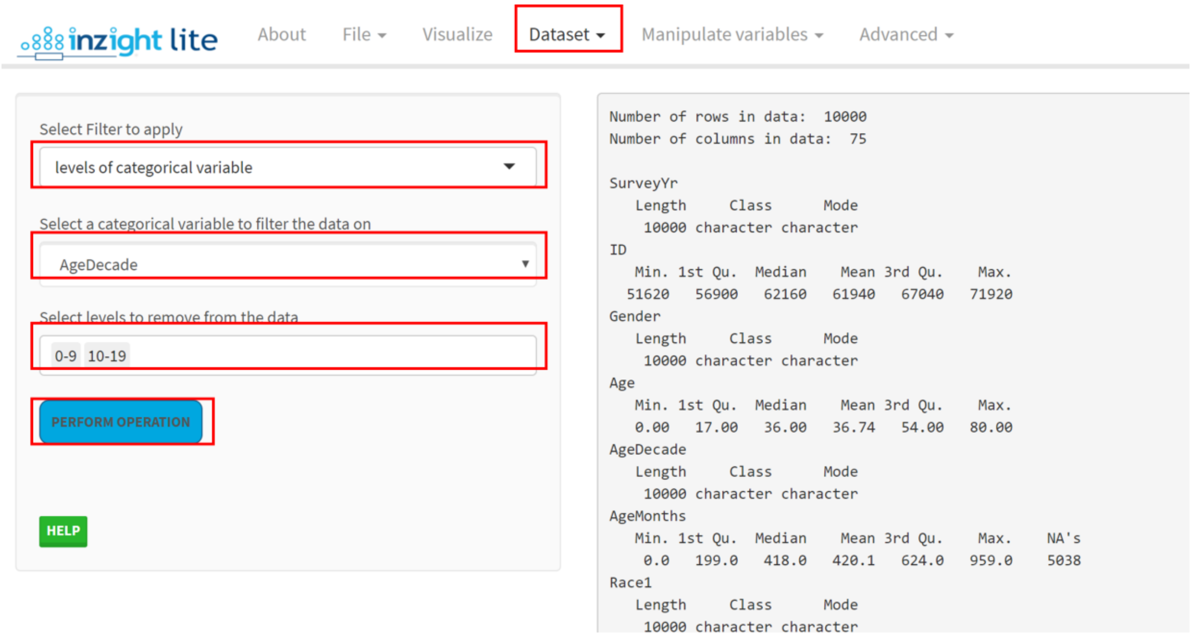
We hope it has been bothering you all this time that there are a large number of missing values for these plots. Why? If you look at the documentation for the data sets, Education is not recorded for those aged under 20.

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, select

> **Dataset** > **Filter Dataset**

- Select Filter to apply: **levels of a categorical variable**.
- Select the variable you wish to filter, e.g. **AgeDecade**
- Select ALL of the categories you wish to remove from the drop down, e.g. **0-9 and 10-19**
- Click on **Perform Operation**



Select Filter to apply: levels of categorical variable

Select a categorical variable to filter the data on: AgeDecade

Select levels to remove from the data: 0-9 10-19

```
Number of rows in data:  10000
Number of columns in data:  75

SurveyYr
   Length     Class      Mode 
    10000 character character 
ID
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  51620   56900   62160   61940   67040   71920 
Gender
   Length     Class      Mode 
    10000 character character 
Age
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00   17.00   36.00   36.74   54.00   80.00 
AgeDecade
   Length     Class      Mode 
    10000 character character 
AgeMonths
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
    0.0   199.0   418.0   420.1   624.0   959.0    5038 
Race1
   Length     Class      Mode 
    10000 character character 
```

The dataset will now include only the remaining age groups and you can recreate the graph for **Education** 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 345, almost all of them have AgeDecade missing.

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 on the discussion.

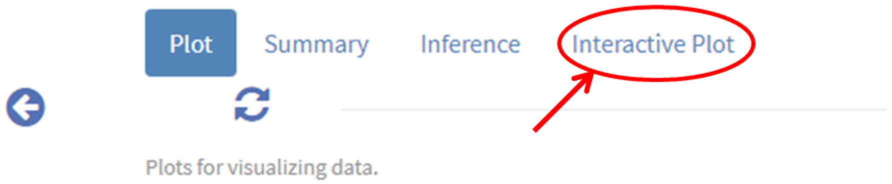
Now create side by side graphs for each level of **Education.** Clear **AgeDecade** from the **subset by** slot, then select **AgeDecade** in the ***second* variable** slot.

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 **subset by** slot. Practise moving the slider between graphs.

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 bars or clicking them. Explore!

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.

Other ways of representing relationships between 2 categorical variables

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?

- Go to **Dataset** > **Restore Data** and then **click** the **RESTORE DATA** button.