MMA 860 – Class 1 Exercise: Data Exploration

The smoking data file (Youth Smoking Survey, 2010-2011) is on the portal as is the data dictionary. Please work with the people next to you if you are having trouble.

Instructions

Working with larger data files:

1. **Download the necessary files from the portal**
    - The smoking data 2000 rows file
    - Data dictionary

2. **Produce a histogram** of slst30a1
   i.e. the number of days last month that someone smoked one or more cigarettes.

3. **Examine the results**
   What appears to be wrong with this result?

   Check the description of slst30a1 in the data dictionary. You will see that the data contains exception codes and is not actually the number of days but a series of categories.

4. **Create a query to filter to only the 'good' records.**
   (Hint: remove the rows with the 'special value' that provides no extra info).

5. **Use the data dictionary to change the categorical data (1-8) to the average number of days that person smoked the last month.**
   You will need to use the label column in the data dictionary for this.

6. **Redraw the histogram**