For my first capstone project, I am using a data set from the UC-Irvine Machine Learning Repository to classify individuals into two categories, above $50k salary and under $50k salary, based on census information. The data set is cleaned for the most part already. Missing values are indicated by a ‘?’. For the initial portion of the Capstone project, I will simply remove data samples where there are missing values. An additional extension of the project could be to attempt imputation on the missing data values.

The next step of the cleaning process is to remove unnecessary data columns and combine similar categorical variables. Firstly, I will eliminate the fnlwgt and education-num columns. The fnlwgt column is a weighted estimate of the population used by the Census Bureau that is not applicable to the classification problem. Education-num, or the years of education an individual has, is redundant with the education level column. Alternatively, the education level column can be deleted to treat education as a numerical variable rather than a categorical one.

Next, similar categories for each column are condensed to reduce the number of categories and simplify the classification process later on. These combinations will be vetted to make sure they have similar percentages of >$50k income.

**Workclass**: Self-employed (self-emp-not-inc, self-emp-inc), Not paid (without-pay, never-worked)

**Education**: Below HS (12th, 11th, 10th, 9th, 7th-8th, 5th-6th, 1st-4th, Preschool), HS Grad (HS-Grad, some college), Associates (Assoc-acdm, Assoc-voc)

**Marital-Status**: Not married (divorced, separated, married-spouse-absent)