Stat 333 Project: Exploratory Data Analysis

Larry Hernandez

February 18, 2016

This project focuses on the relationship between 4-year adjusted high school graduation rates in 2013 and median household income (2013) by school district.

Exploratory data analysis reveals that these data need cleaning. About 66% of the graduation rates data are not reported as a single number. For California and Wisconsin, 30% and 75% of the graduation rates are reported as something other than a number (See Barchart). Inspection of these data using the 'View' function reveals that many graduation rates are reported as ranges (i.e. 80-84) or with some other notation, such as "GE95" or "PS", which is actually a code to indicate that the privacy of the students was maintained via reporting this code in lieu of an actual numerical graduation rate. "GE95" might actually denote a 95% graduation rate; this is verifiable by visiting school district websites & gathering published graduation numbers. Other alphanumeric values in this dataset might need to further verification or they will simply be omitted.

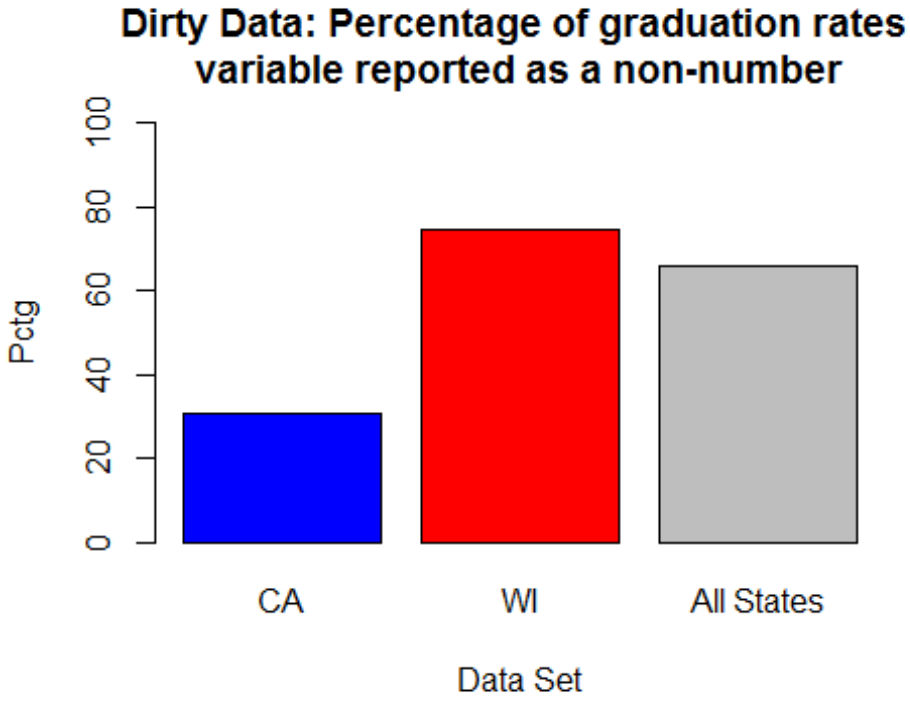
Some variables (ie, Limited English Proficient, various Ethnic categories, Total-reduced price Lunch Eligible Students) from the additional demographic data set (ElSi tableGenerator, NCES) have special symbols for their reported values, such as the dagger (†), double-dagger (‡), or minus sign (-), with meanings such as "not applicable", "Data do not meet NCES standards", and "indicates that data are missing", respectively. Occasionally a value is reported as "n/a" or other variant. Some numerical values for Ethnicity are preceded by an "=" sign and will have to be revised with a text-parsing function if Ethnicity is to be used in this analysis.

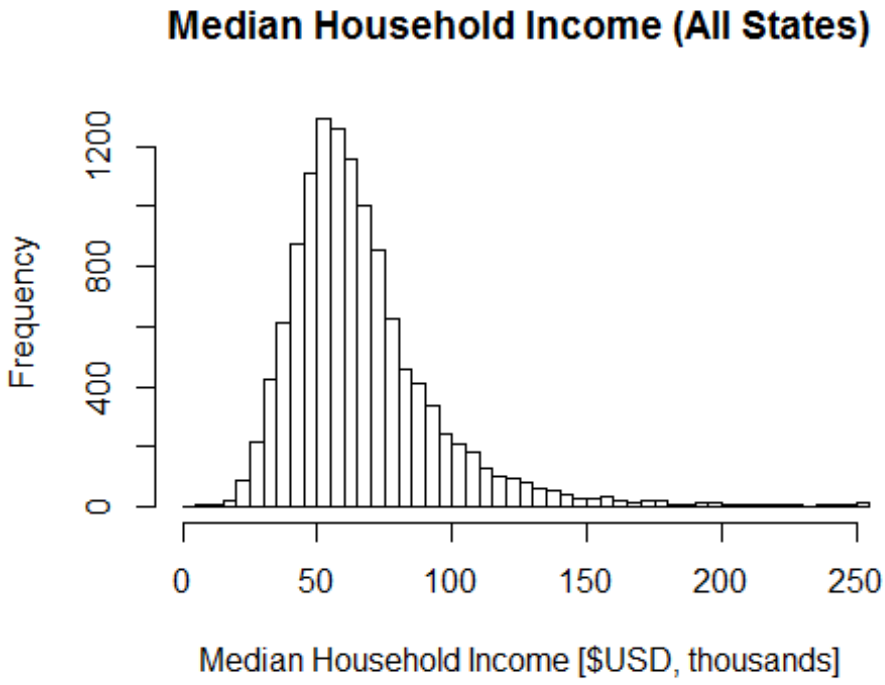
A histogram of the median household income for all districts indicates that the data are centered about $50,000. This is reasonable for 2013, in which median household income was about $52,000.

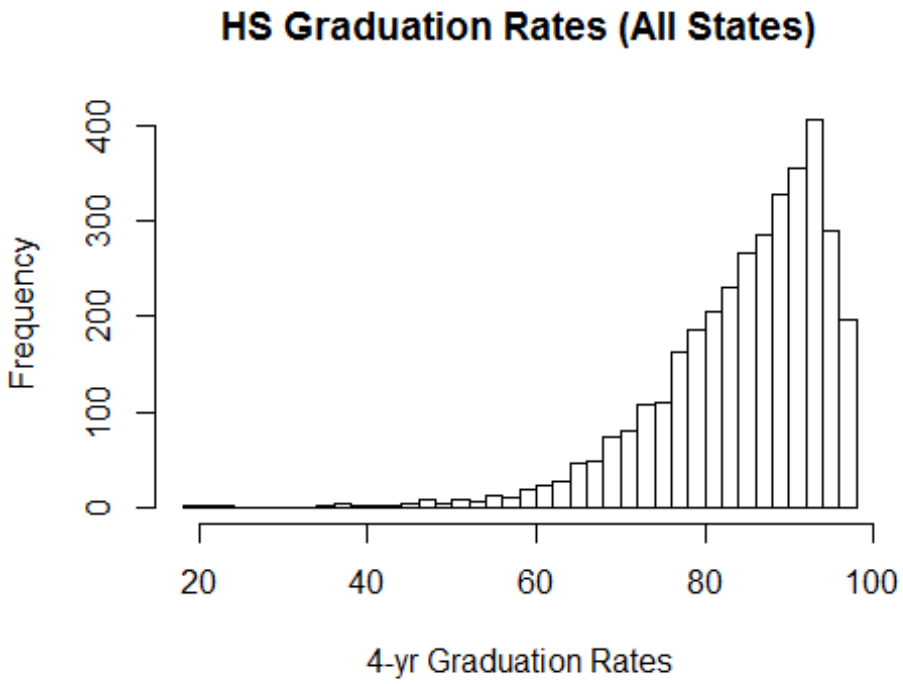
The histogram of the graduation rates (for the 35% which have been reported as a number) does not reveal any obvious problems, such as negative values or values greater than 100.

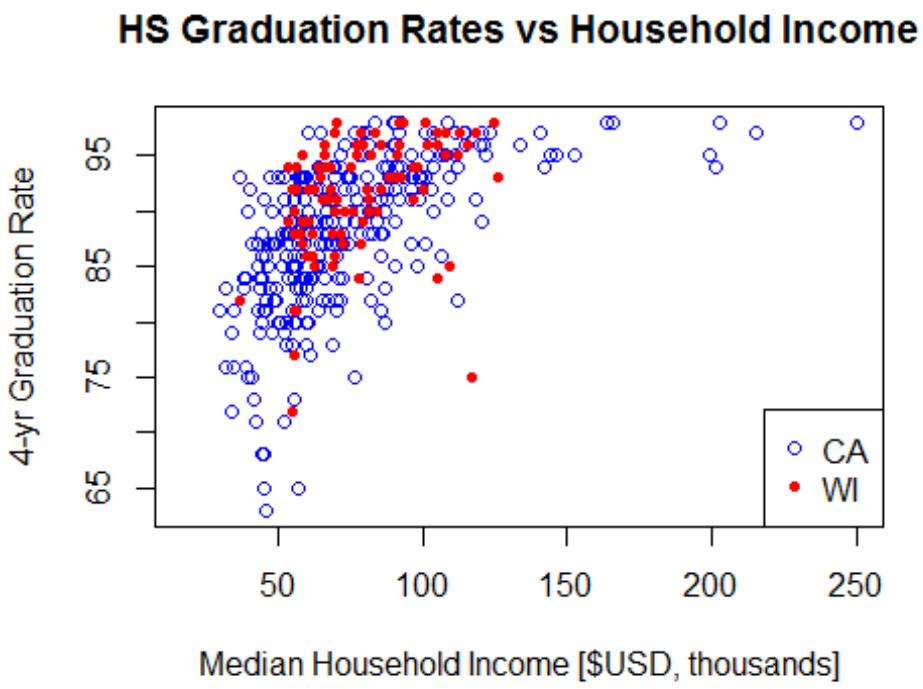
The last two figures are scatterplots of 4-year high school graduation rates vs median household income (in units of thousands of dollars). Both scatterplots indicate that there is a non-linear relationship between the two variables, with very high graduation rates for school districts with median household income above $150k, and a lot more variability around $50k. This relationship appears to follow the same pattern for CA, WI, and the data set that excludes those two states. That is encouraging.

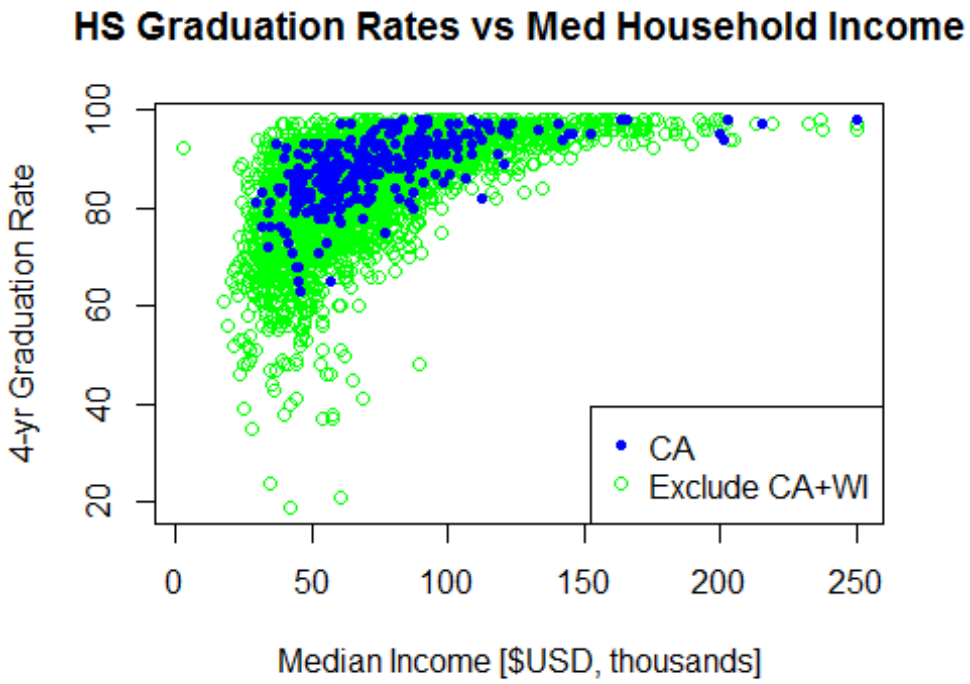
Note: This rmarkdown file was generated with the option "warning=FALSE" to suppress warnings about coercion of NA values.


Dirty Data: Percentage of graduation rates
variable reported as a non-number
Pctg
0
20
40
60
80
100
CA
WI
All States
Data Set


Median Household Income (All States)
Frequency
0
400
800
1200
0
50
100
150
200
250
Median Household Income [$USD, thousands]


HS Graduation Rates (All States)
Frequency
0
100
200
300
400
20
40
60
80
100
4-yr Graduation Rates


HS Graduation Rates vs Household Income
4-yr Graduation Rate
65
75
85
95
50
100
150
200
250
Median Household Income [$USD, thousands]
CA
WI


HS Graduation Rates vs Med Household Income
4-yr Graduation Rate
20
40
60
80
100
0
50
100
150
200
250
Median Income [$USD, thousands]
CA
Exclude CA+WI