PSTAT 126 Project Step 1

Alec Wang, Gary Han

2023-10-22

Introduction

This dataset includes 14 parameters from 30162 adults collected during the 1994 census as predictors for whether or not income exceeds $50k/yr. We have sampled 500 entries from the dataset for analysis. Below is a table detailing the 14 parameters and the response, that were collected in the census.

Field	Description
age	Age in years of individual (Integer)
workclass	Class of work of individual (7 categories)
fnlwgt	Number of people the entry represents (Integer)
education	Highest level of education of individual (16 categories)
education-num	Maps each category in education to a number (Integer)
marital-status	Marital status of individual (7 categories)
occupation	Description of occupation (14 categories)
relationship	Relationship of individual relative to others (6 categories)
race	Category of race of individual (5 categories)
sex	Biological sex of individual (2 categories)
capital-gain	Capital gain of individual (Integer)
capital-loss	Capital loss of individual (Integer)
hours-per-week	Hours worked per week by individual (Integer)
income	Whether or not income is above $50k/yr (2 categories)

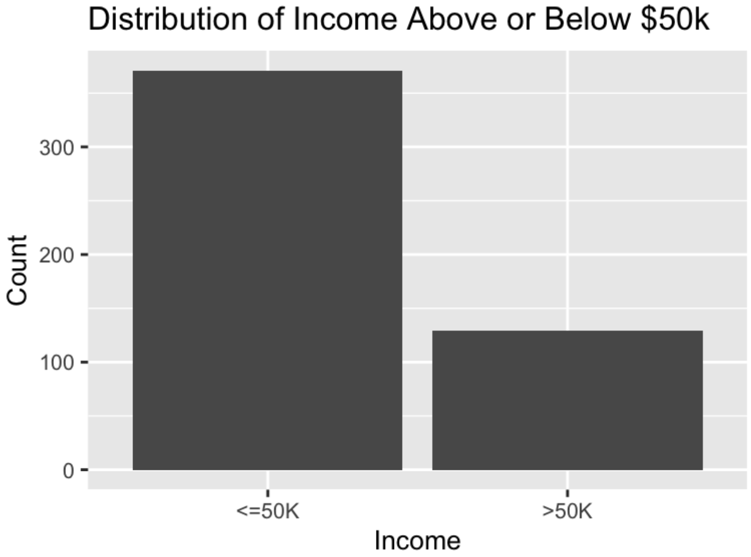
Summary of variables

Summary of Numeric Variables

	Min	Q1	Median	Q3	Max	Means
age	17	27.0	37	47	90	38.418
fnlwgt	22831	116408.5	180952	242053	648223	189666.834
education-num	2	9.0	10	13	16	10.146
capital-gain	0	0.0	0	0	99999	923.208

	Min	Q1	Median	Q3	Max	Means
capital-loss	0	0.0	0	0	2282	60.682
hours-per-week	8	40.0	40	45	98	40.900

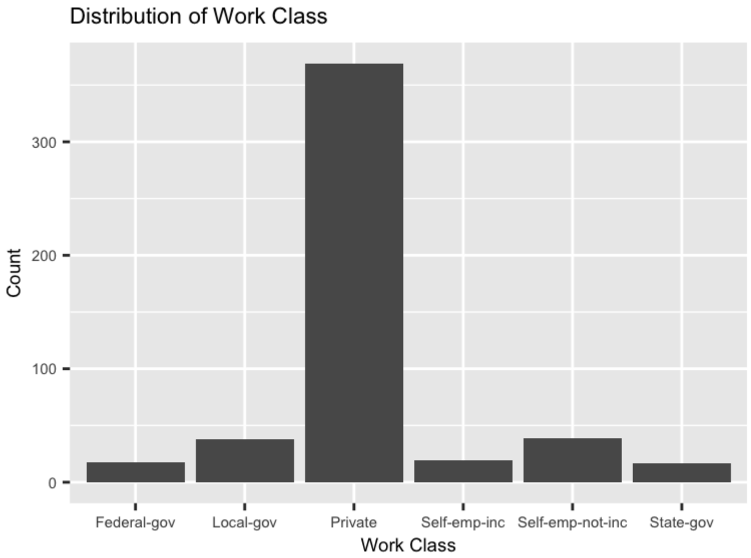
We notice that both the capital gain and capital loss data are significantly skewed right, as at least 75 percent of all entries in either set is 0. Interestingly the maximum values for capital gain seems to be capped below 100,000, which may suggest that there is some upper limit to capital gain that an individual may report.



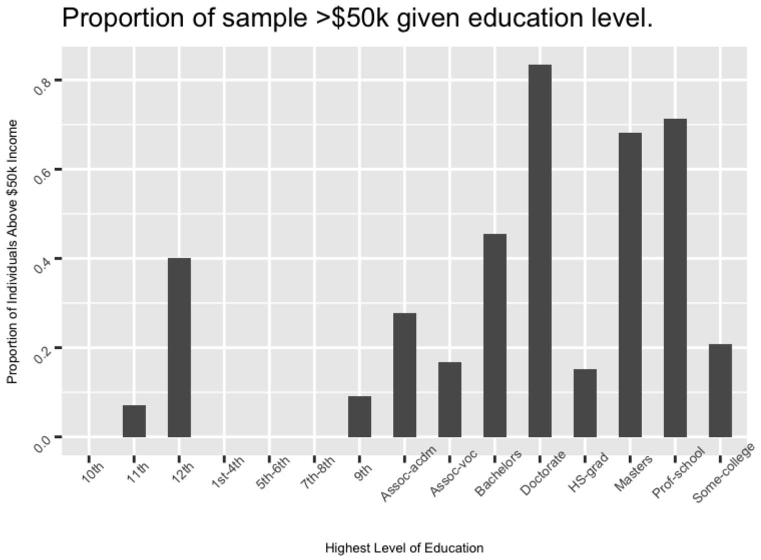
Number of people in each of the two income brackets. We can notice that only about 25% of the sample has an income greater than $50k.

Examining Relationships

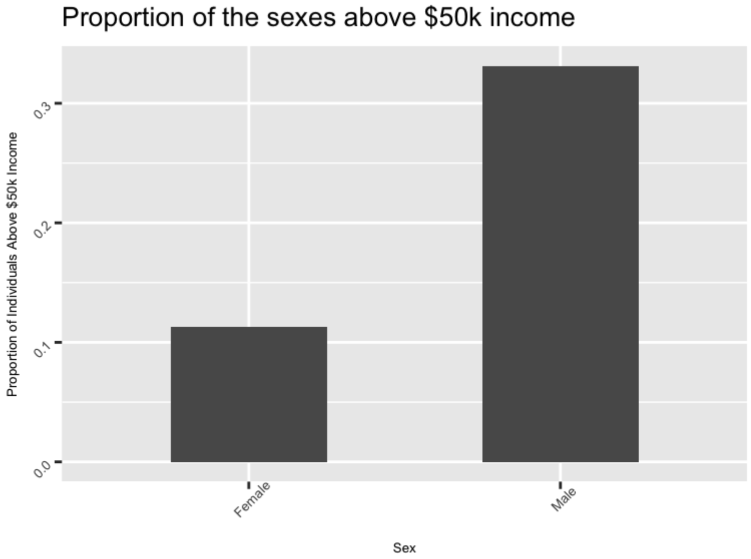
In the following graphs, we will examine the relationship between income and some of the categorical variables.



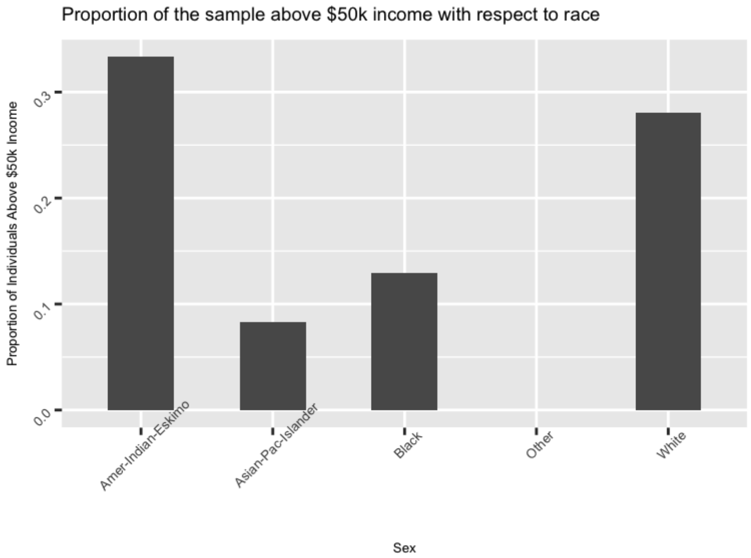
Number of people in each work class. We can see that the vast majority of sampled individuals work as employees in non-government companies. This is to be expected in a capitalist society since most people are working for each other instead of the government.



Proportion of individuals earning above $50k per year within each level of education. We noticed that there is a positive correlation between level of education and income, with the doctoral and professional school graduates having the highest proportion earning above $50k.



Proportion of individuals above $50k income based on sex. There is a significantly greater proportion of males with income above $50k, which suggests that males would earn more on average (i.e. if you were a male you would have a 30% chance of making more than 50k, whereas if you were female, that chance drops to a little over 10%).



Proportion of the sample above $50k income based on race. We noticed that there was a significant proportion of American Indians and Eskimos with an income above $50k.

Concluding comments

The data is about what we expected, except for the statistics on income based on race. Based on data from a census published in 2001, we would have expected the income for Native Americans to be lower and the income for Asian/Pacific Islanders to be much higher. We think this discrepancy is due to either sampling variance, heavily right-skewed distributions, or the census not properly representing the population. Another possibility is that this is how the data from 1994 is actually distributed, and the distributions shifted by a significant amount in the span of 7 years.

The sampling of the data went well. It was the plotting that took the most amount of time, most of which was spent trying to debug `ggplot`. For the most part, we think that we had a representative sample of the population, since we took the sample out of the US census, which itself tries to be as objective as possible when gathering information.