The file ipums.csv contains data extracted from the 1980 census Integrated Public Use Microdata Series (IPUMS). See the file ipums.pdf for the descriptions of all variables. **We treat these data as a population.**

The variable inctot is total personal income from all sources. Note that it is "top-coded" at $75,000 to protect the confidentiality of the respondents. Nevertheless, ignore the effect that the top-coding will have on estimates.

PART ONE: CLUSTER SAMPLING

(1A) Generate a frequency table of the number of persons within each psu.

(1B) Take a two-stage cluster sample of size 200 as follows: first select an SRS of 10 psus and then in each of the selected psus take a subsample (SRS) of persons with sample size proportional to number of persons within the psu. (You might want to use the R function 'set.seed' so that you can reproduce the sample.)

(1C) Using your sample, estimate the population mean and total of 'inctot' and give the standard errors of your estimates.

(1D) Select an unequal-probability sample of 10 psus, with probability proportional to number of persons within each psu, and then take a subsample (SRS) of 20 persons in each of the selected psus. (You might want to use the R function 'set.seed' so that you can reproduce the sample.)

(1E) Using your sample, estimate the population mean and total of 'inctot' and give the standard errors of your estimates.

PART TWO: COMPLEX SAMPLING

(2A) Select a two-stage stratified cluster sample as follows: first select two psus from each stratum, with probability proportional to size, and then take a simple random sample of 50 persons from each selected psu. (You might want to use the R function 'set.seed' so that you can reproduce the sample.)

(2B) Construct the column of sampling weights for your sample.

(2C) Draw a histogram of 'inctot' for your sample that incorporates weights.

(2D) Draw two of the scatterplots that incorporate weights for y variable 'inctot' and x variable 'age'. How do these differ from scatterplots that do not use the weights?

(2E) Using your sample, estimate the population mean and total of 'inctot' and give the standard errors of your estimates.

(2F) Create a categorical variable catinc from 'inctot' with two categories: low income and high income, using the median income as the dividing point for the categories.

(2G) Using your sample, conduct a hypothesis test to explore whether 'catinc' is associated with 'race'. What method did you use to account for the complex sampling design?

WRITE UP A REPORT CONTAINING YOUR FINDINGS AND/OR CONCLUSIONS.