Problem Set #1

MACS 30000, Dr. Evans
Xinzhu Sun(12147991)

Problem 1 Write a data section for your assigned data set.
Part (a). The data can be accessed from http://thedataweb.rm.census.gov/ftp/cps_ftp.html#cpsbasic . It's stored and curated jointly by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics (BLS).

Part (b). Papers that used the CPS data of Jan. 2016 cannot be found for the data set is too new to be use, even the BLS own economists team published no working paper that has Jan. 2016 CPS data included. However, key papers that used the past CPS data are listed as following:
1. Madrian, B. and L. J. Lefgren, An Approach to Longitudinally Matching Current Population Survey (CPS) Respondents, *Journal of Economic and Social Measurement*, 2000, 26: 31-62.
2. Barry T. Hirsch and David A. Macpherson, Union Membership and Coverage Database from the Current Population Survey: Note, *Industrial and Labor Relations Review, Vol. 56, No. 2*, January 2003, pp. 349-54.
3. Franco Peracchi and Finis Welch, How representative are matched cross-sections? Evidence from the Current Population Survey, *Journal of Econometrics,Vol. 68, Issue 1*, July 1995, pp. 153-179.
4. Jin Heum Park, Estimation of sheepskin effects using the old and the new measures of educational attainment in the Current Population Survey, *Economics Letters, Vol. 62, Issue 2,* 1 February 1999, Pages 237?240.
5. Richard V. Burkhauser, Kenneth A. Couch and David C. Wittenburg, A Reassessment of the New Economics of the Minimum Wage Literature with Monthly Data from the Current Population Survey, *Journal of Labor Economics, Vol.18, No.4*, October 2000.
6. Donald R. Shopland, Anne M. Hartman, James T. Gibson, Michael D. Mueller, Larry G. Kessler and William R. Lynn, Cigarette Smoking Among U.S. Adults by State and Region: Estimates From the Current Population Survey, *Journal of the National Cancer Institute, vol.88, Issue 23*, 4 December 1996, 1748-1758.
7. Unpublished BLS working paper can be found from the link: https://www.bls.gov/osmr/home.htm#publications.
etc.

Part (c). The data is collected by direct surveys of samples of the population.

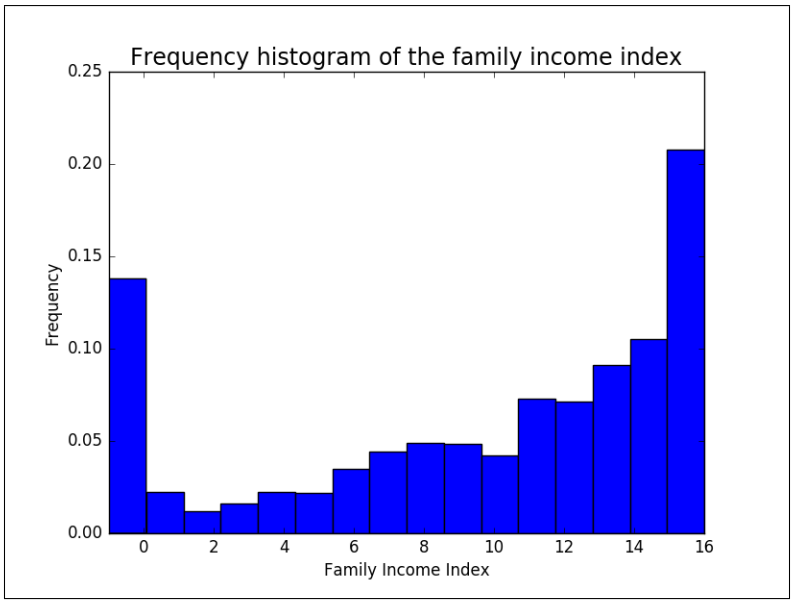
Part (d). Considering the fact that almost all variables are presented in index form in this data, it's meaningless to calculate the mean and standard deviation of those variables. In this case, I choose quantiles and mode as my descriptive statistics.

Table 1: Variables from CPS

Variable	Description	Meadian	75% Quantile	Mode
HEHOUSUT	TYPE OF HOUSING UNIT 1: HOUSE, APARTMENT, FLAT	1.0	1.0	1
HEFAMINC	FAMILY INCOME 6: 15,000 TO 19,999 11: 40,000 TO 49,999 14: 75,000 TO 99,999	11.0	14.0	NaN
HURNUMHOU	TOTAL NUMBER OF PERSONS LIVING IN THE HOUSEHOLD	3.0	4.0	2
GEREG	REGION 2: MIDWEST (NORT CENTRAL) 3: SOUTH; 4: WEST	3.0	4.0	3
PEMARITL	MARITAL STATUS 1: MARRIED - SPOUSE PRESENT 4: DIVORCED	1.0	4.0	1
PEEDUCA	HIGHEST LEVEL OF SCHOOL 39: HIGH SCHOOL GRAD(GED) 41: ASSOCIATE DEGREE	39.0	41.0	NaN
PTDTRACE	RACE 1: White Only	1.0	1.0	1
PRMARSTA	MARITAL STATUS (ARMED FORCES PARTICIPATION) 1: MARRIED, CIVILIAN SPOUSE PRESENT 5: DIVORCED	1.0	5.0	1

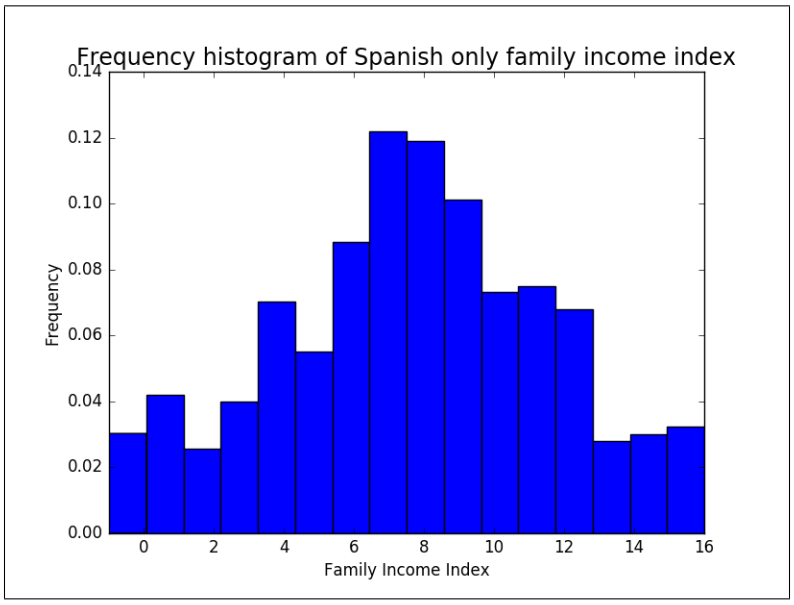
NaNs raise from two values have same number of counts.

Part (e). The histogram of distribution of family income type is as below: Sur-



prisingly, it can be seen from the plot that the CPS tend to include people with higher income or extremely poor people. It seems the CPS tend to speak for wealthy people.

Part (f). The histogram of distribution of Spanish only family income type is as below: We find another interesting fact that in those Households that Spanish is the



only language spoken by all members of this household who are 15 years of age or older, the income types are distributed almost normally. Thus I guess the wired plot in part (e) may raise from sampling among different race or ethnicity.