Name: _____ Math 227 / Fall 2019 / Prof. David Soto

Module 3. Activity 2.

Question 1. I have uploaded a data file of 2017 personal income from a random sample from the census bureau. Use the following instruction to access the data and save it to a data frame **PersonalIncome:**

PersonalIncome <- read.csv("https://tinyurl.com/y5aogpps", header=TRUE)

How many observations do you have in this data frame? Which instruction did you use to verify this?

Question 2. Create another data frame that contains only adults (18 year-old or older). Call this data frame **PersonalIncomeAdults**.

What instruction did you use to do this? How many observations do you have in this data frame?

Question 3. We might have some "unwanted" data under **EMPSTAT.** See the variables codes given at the end; which records would you discard? Why?

Question 4. Clean **PersonalIncomeAdults** to discard the codes you decided in Question 4. Save the new data under **PersonalIncomeAdults2.** (Note: in R the logical statement "x or y" is written " x|y").

What instruction did you use to do this? How many observations do you have in this data frame?

Question 5.

Create a histogram for the variable **INCTOT** of the data frame **PersonalIncomeAdults2.** Use a binwidth of 50,000. What is the income bracket with the largest number of individuals? About how many individuals do you find in this income bracket?

Question 6.

The income bracket at the bottom end of the distribution extends from $-25{,}000$ to $25{,}000$. About which percent of the total number of observations fall in this income bracket? Based on this, would you expect the first quartile of the total income (**INCTOT**) to be more or less than $25,000? Explain.

Question 7.

About which percent of the total number of observations fall in the income bracket extending from 25,000 to 75,000 dollars? Based on this, would you expect the third quartile of the total income (**INCTOT**) to be more or less than $75,000? Explain.

Question 8. What is the five-number summary of total income?

Question 9. What is the income level that separates the high outliers? Show your calculation. Would you discard the high outlier in an economic analysis about income inequality? Explain.

SEX:

Value	Label
1	Male
2	Female

RACE :

Value	Label
1	White
2	Black/African American/Negro
3	American Indian or Alaska Native
4	Chinese
5	Japanese
6	Other Asian or Pacific Islander
7	Other race, nec
8	Two major races
9	Three or more major races

EMPSTAT

Value	Label
0	N/A
1	Employed
2	Unemployed
3	Not in labor force

INCTOT Specific Variable Codes
-009995 = -$9,900 (1980)
-000001 = Net loss (1950)
0000000 = None
0000001 = $1 or break even (2000, 2005-onward ACS and PRCS)
9999999 = N/A