**R Practice with a relatively big data set.**

Today, we will build on the R skills you’ve learned in the past and/or during the last class by using the real-life survey data set and arriving at some stumbling blocks that real datasets can present. Then, we will start using R to analyze the data and come to some conclusions.

**Case:** You are a real estate development analyst aiming to assist developers in improving access to low-income housing to low-income families. You have been given a dataset by the U.S. Department of Housing and Urban Development that contains survey data for 64,535 low-income households in the U.S. This survey is representative of the U.S. low-income households.

**Descriptives**

1. What is the mean, median, and standard deviation of household salary (Variable:totsal) in the sample? What issue did you come up with when you try to calculate these in R? How did you solve the issue?
2. Create a histogram of household salary. What stands out to you?
3. How many (and percentage of) households had 0 dollars of salary? One way to easily do this is to generate a new binary variable of 1’s and 0’s (where 1 = missing and 0 = not missing), then find the mean of that variable. A useful link to help (see part on recoding variables): <https://www.statmethods.net/management/variables.html>
4. How many (and percentage of) households has missing salaries? What do you think is the indicator for missing salaries in this data? Why do you think there are missing salaries? Why do you think the indicator for missing salaries is what it is.
5. Create a scatterplot showing head of household age (‘age’) on the X-Axis and salary (‘totsal’) on the Y-Axis. Add a regression line. Calculate the correlation between the two variables. Then, run a simple linear regression. What can you tell about the relationship between the two variables?

**Additional Analyses**

Housing cost burden is calculated as the annual housing cost divided by the total salary. Annual housing cost is given by “cost08” and total salary is “totsal.”

1. Calculate the average cost burden for the sample. Also, calculate the standard deviation.
2. Also calculate the 25th, median, and 75th percentile of the burden.
3. Do whatever univariate analysis on burden that you would like. What are some things that stand out about the burden of housing cost on these families?
4. The recommended housing burden is generally 30%. What percentage of families in the sample has a burden of greater than the recommended amount? What happens to burden when you add utilities (‘utility’) and other related housing costs (‘other costs’) to calculate the burden?
5. Is the burden higher or lower for older head of households? How do you know?
6. Your boss knows the survey participants are representative of low-income families in the U.S. In terms of your analysis on burden, are there any potential errors that we must consider that would affect whether the sample statistic is representative of the population parameter?

**Big Question**

Your boss has built a very successful business where you match low-income households to housing that they can afford and is very passionate about her business. As a part of successful family-housing matches, the business charges a small fee to the family. After taking a look at the data on housing burden, she becomes distraught that her business produces more burden to her clients. After taking a look at the descriptive statistics and doing some analysis on burden, what are some strategic ways that you could advise her to alleviate household burden while remaining profitable?

Also, your company only operates in the East Region. Would this change your analysis? How should your boss feel now with the new analysis? Think about what errors/biases there might be also with this analysis.