

Activity 14: Using large scale datasets

The Current Population Survey (CPS) is “one of the oldest, largest, and most well-recognized surveys in the United States.” It is immensely important, providing information on many of the things that define us as individuals and as a society – our work, our earnings, and our education. Students can access raw data as well as details of methodology. Students will need to prepare the data for analysis, including consulting the Data Dictionary for information about the variables, and engage in data cleaning.

(a) Open the CPS webpage <https://www.census.gov/programs-surveys/cps.html>. Follow the links for Technical Documentation and then Methodology. Provide a brief summary of what you learn about how these data are collected (e.g., How often is the survey conducted? How many people? What are the observational units? Do they use random sampling? Are any individuals excluded from the data? What is the idea behind “weighting”? Who stands to benefit from these data? Does anyone stand to be harmed by this data?)

An example of a quick opportunity to engage students in the data collection process, methodology decisions, and ethical considerations. Most students can ‘critique’ these approaches from the beginning of the course, as well as ask questions they want to know more about.

(b) Open the Data page <https://www.census.gov/programs-surveys/cps/data.html> and follow the third link for Current Population Survey Datasets. Follow the link for Annual Social and Economic Supplements. Download the CSV file from under Data and Documents. This downloads a zip file. Extract the files, you want the pppub22.csv file. How many observations are in the data file?

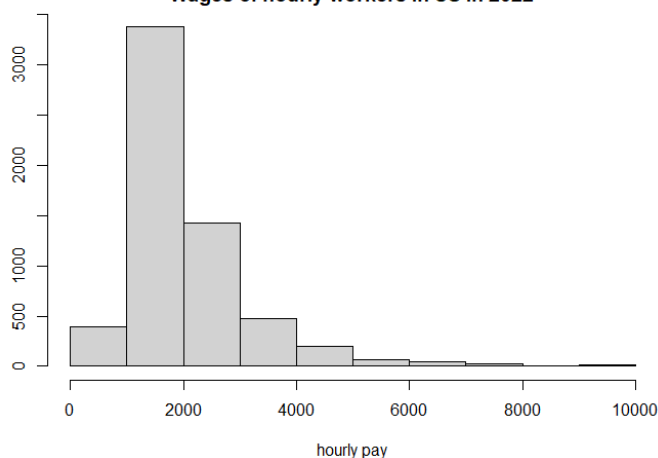
Good practice to have intermediate checks along the way to make sure everyone is using the correct data.

(c) Back on the CPS webpage, open the [Data Dictionary](#). Find the description of the A_HRSPAY variable. What does this variable measure? Who is measured for this variable? ALSO, back in the Data Dictionary, which variable reports the biological sex of the individual? (How did you find it?)

Good practice to show students tools for finding answers to their questions (e.g., data dictionary) and strategies for making the data less overwhelming (e.g., ctrl F).

First, have students explore the data, e.g., with a dotplot. This is a good example of where a dotplot is not very useful (because there are so many observations) and also highlights some of the utility of histograms as well as data summaries. They will soon see a stack of observations at 9999.

Wages of hourly workers in US in 2022



Consulting the data dictionary, they will see that 9999 is how the missing values/non respondents/not in universe responses to this likely 4-digit question are coded. Students should quickly recognize the need to remove these observations from the dataset before performing any analysis.

(d) In the data file, subset the data to only include the individuals with A_HRLYWK = 1 and A_HRSPAY > 0. How many individuals remain?

Provide students with lots of hints, even video demonstrations. Being able to subset data is a valuable skill for data cleaning and exploration.

(e) Make a well-labeled histogram of the A_HRSPAY variable for the new subsetted data, also report the mean, standard deviation, and five number summary (with measurement units, make sure your results make sense in context!). Write a one-sentence summary of the behavior of this distribution in context.

Use the rich dataset to apply skills from the course.

(f) Describe what the shape of this distribution implies in this context. Is the comparison between the mean and median hourly wages consistent with this shape? Explain.

Income type data is typically skewed to the right. I often follow this example up with a graph from the media showing how the mean and median have grown at different rates over time illustrating an increasing skewness to the distribution as “the rich get richer.”

(g) Determine the median hourly wage for women and median hourly wage for men.

(h) Examine the ratio: how much do women make for every \$1 men make, “on average.”

Students can choose between comparing means or medians! You can tie this number to recent news reports. Students seem to appreciate understanding where this number comes from/creating it themselves.

(i) Which Race(s) are NOT present in this dataset? (Hint: Data Dictionary)

This is again mostly just to further engage them in the data, get them comfortable with large datasets, think about why you might want to reduce the number of categories, consider which categories are often ignored, how many people don't answer some questions etc.

(j) Find the median hourly wages for the following categories. Document your steps. (This one is a bit of a technology challenge question.)

White (only) male	
White (only) female	
Black (only) male	
Black (only) female	
Asian (only) male	
Asian (only) female	

(k) Summarize what you learn from these six values. Also, suggest 1-2 additional variables (quantitative or categorical) that you think might explain some of these differences in wages.