Data Science for Public Policy

Aaron R. Williams - Georgetown University

# PPOL 670 | Assignment 3
# Applied Introduction to R's Tidyverse

**Due Date:** Friday, February 11 at 6:00 PM

**Deliverable:** A single, well-documented `.pdf` created with R Markdown. Each exercise should be clearly labeled.

## Rubric

- Submit a .pdf created with the .Rmd template.
- Each exercise is clearly labeled with Markdown.
- Attempted exercises will receive more than zero points. Unattempted exercises will receive zero points.

**Points:** 10 points

### Overview

The U.S. federal government is a rich source of data for research including large scale surveys like the American Community Survey (ACS) and Current Population Survey (CPS). Data from surveys are typically released as published tabulations and are sometimes released as microdata files with information about many variables for each sampled unit (person, household, etc.). Microdata can be used to answer a much wider range of questions than publsihed tabulations, but microdata sometimes aren't released because of disclosure concerns.

This assignment explores published microdata from the CPS with tools from class like `library(dplyr)` and `library(ggplot2)`. This assignment also introduces `library(srvyr)` and demonstrates how to calculate summary statistics from data gathered from probability samples.

## Project set up (1 point for the .pdf)

1. Create a new folder on your computer called `assignment03`.
2. Open R Studio.
3. In the top right, click "Project: (None)".
4. Click "New Project".
5. Create a new project in the existing directory `assignment03/`.
6. Copy-and-paste `template.Rmd` into `assignment03/` and rename it `assignment03_<NetID>.Rmd`
7. Edit the `author` field in the YAML header at the top of the file to your name and NetID.

8. Click knit to ensure that you have necessary installations. You will likely need to install `library(ipumsr)` and `library(srvyr)`.

# Exercise 01 (1 point)

### About the Data

We will access the CPS from IPUMS CPS. The CPS is a monthly household survey for the Bureau of Labor Statistics gathered by the US Census Bureau. The questionnaire asks many important questions about demographics and labor conditions. The CPS is the data source for the official unemployment rate in the United States. In addition to core questions, the CPS often contains supplements like the Annual Social and Economic (ASEC) supplement, which is also known as the "March supplement".

Integrated Public Use Microdata Series (IPUMS) are created by the Minnesota Population Center. IPUMS harmonize major US data sets over time and increase the usability of the data by standardizing files, variables, codes, and documentation.

You can read more about the CPS and IPUMS CPS here.

1. Register for IPUMS CPS by selecting "Apply for Access" if you do not have an IPUMS account or "Login" if you do have an IPUMS account and following the instructions.
2. Navigate to the IPUMS homepage and select "Get Data" under "CREATE AN EXTRACT".

We need to select **samples** (observations) and **variables**.

3. Click "SELECT SAMPLES", unselect "All Default samples", and then select April 2020 and April 2021 on the "Basic Monthly" tab.
4. Select "PERSON", "CORE", and "DEMOGRAPHICS". Then add the following variables to the cart: `RELATE`, `AGE`, `SEX`, `RACE`, `MARST`, `POPSTAT`.
5. Select "PERSON", "CORE", "WORK". Then add the following variables to the cart: `EMPSTAT`, `LABFORCE`, `OCC`, `IND`, `UHRSWORKT`, `UHRSWORK1`, `UHRSWORK2`, `AHRSWORKT`, `AHRSWORK1`, `AHRSWORK2`, `ABSENT`, and `WKSTAT`.
6. Click "VIEW CART", review your selections, and click "CREATE DATA EXTRACT". Add a description that includes the class and assignment number. Click "SUBMIT EXTRACT".

While IPUMS processes the data request, we can download a few important files.

7. Right click on R and Save Link As to your assignment directory.
8. Right click on DDI and Save Link As to your assignment directory. The file should be an .xml file.
9. Once you have received your email from IPUMS, follow the link and "DOWNLOAD .DAT" into your assignment folder.
10. Install `library(ipumsr)`. Load `library(tidyverse)` and `library(ipumsr)` in your .Rmd script.
11. Copy-and-paste lines 6 and 7 from the .R script in step 7 into your .Rmd and load the data.
12. Add `glimpse(data)`.

# Exercise 02 (2 points)

The CPS is a complex survey and unweighted statistics will not represent the population of interest (they are incorrect). We must account for the survey design and weights when calculating statistics. `library(srvyr)`, which is a tidyverse-friendly extension of `library(survey)`, provides tools for working with complex surveys.

There are two main steps to calculating statistics from complex surveys with `library(srvyr)`. First, create a `tbl_svy` with `as_survey_design()`. Second, use `summarize()` and special functions like `survey_mean()`, `survey_total()`, and `survey_prop()` to calculate statistics. Optionally, use `group_by()` before `summarize()`.

1. Read this brief introduction to srvyr.
2. Load `library(srvyr)` and use `as_survey_design()` to create a `tbl_svy` object called `cps_svy`. Use `WTFINL` for `weight`.
3. Use `class()` on `data` and `cps_svy`. Note the difference after running `as_survey_design()`.

IPUMS CPS provides clear documentation about variables in the CPS. To see the documentation, navigate to variables on the IPUMS CPS website in the same way you added variables to the cart, clicking on the variable name. The page should have section for CODES, DESCRIPTION, COMPARABILITY, UNIVERSE, and more.

4. Read this page and the documentation for `UHRSWORKT` and `AHRSWORKT`. What is the universe of the question? Is there a value for `NIU`? `library(srvyr)` ruins `count()`. A workaround is to use `unweighted()`. Use code similar to the following code to calculate the number of observation in each year (separately for 2020 and 2021) that are not in the universe for `UHRSWORKT` and `AHRSWORKT`.

```
group_by(UHRSWORKT) %>%
summarize(
  n = unweighted(n())
)
```

5. Filter out observations not in the universe for `UHRSWORKT` and create a new data frame called `cps_subset_svy` with the assignment operator.
6. Using the tip from 2.4, count the number of unweighted responses for each value for `UHRSWORKT`. Then create a bar plot (`geom_col()`) excluding values with "hours vary". You should see bunching at round numbers because respondents often report rounded or approximate amounts (Here's a reference from the 1970 CPS!).
7. Calculate the mean usual hours worked (`UHRSWORKT`) in 2020 and 2021. Exclude workers with "Hours vary".
8. Calculate the proportion of workers who usually worked (`UHRSWORKT`) exactly 40 hours in 2021. To do this, either us `mutate()` to create an indicator variable where 0 = doesn't work 40 hours and 1 = works 40 hours and use `survey_mean()` OR `group_by()` and use `survey_prop()`. Include workers with "Hours vary" as not working 40 hours.
9. Calculate the proportion of workers who worked less, the same, and more than usual in April 2020 and April 2021 (separately) by comparing `UHRSWORKT` and `AHRSWORKT`. You will need to use `case_when()` to create a variable with "work less", "work the same", and "work more". Exclude "hours vary".

## Exercise 03 (2 points)

Federal Reserve Economic Data (FRED) is a data tool created by the Federal Reserve Bank of St. Louis. It contains published economic data from a range of sources including tabulations generated from the CPS.

When possible, it is good to compare tabulations from microdata against published tables. This is one way to check that weights are correctly specified and that code is correct.

1. Review the documentation for `EMPSTAT` and `LABFORCE`.
2. Using `cps_svy`, calculate the weighted count of observations ages 16 or older for each year with `survey_count()`, which should work with `group_by()`. Compare your tabulation with series CNP16OV here. Your results should match the April 2020 and April 2021 values to the closest thousand because of rounding.
3. Use `mutate()` to create three numeric indicator variables (equal to 1 if true, 0 if false):
   - `labor_force` if the observation is in the labor force.
   - `employed` if the observation is employed.
   - `unemployed` if the observation is unemployed.
4. Use `AGE` and `POPSTAT` to filter to the civilian populations ages 16 or older, group by year, and use `survey_total()` to:

- Calculate the number of people in the labor force with `labor_force`.
- Calculate the number of people employed with `employed`
- Calculate the number of people unemployed with `unemployed`

5. Compare the results you calculated with microdata to the official tabulation for April 2021. Your calculations should be close(ish).

Note: The labor force is the sum of employment and unemployment.

## Exercise 04 (2 points)

Survey statisticians sometimes construct different weights to achieve different goals. In addition to reflecting the probability of a unit from the population being included in the sample, these weights can include adjustments for nonresponse and further adjustments to hit known control totals (e.g. if the weighted count of people ages 16 and over is 0.2% lower in the CPS than the Decennial Census, the statistician may adjust the weights so the counts match the Decennial Census).

The published labor market statistics from the monthly BLS employment situation report use composite weights (this link is way in the weeds and definitely isn't required). Unfortunately, IPUMS CPS does not include the composite weights.

In this exercise, we will read in the composite weights from the National Bureau of Economic Research (NBER), join them to the IPUMS USA data, and then re-calculate the summary statistics from the second half of exercise 03. To download the composite weights, go to the NBER link, click the "In Stata .dta format" link under downloads, download the cpsb202104.dta file, and place the downloaded file in your assignment folder.

`library(haven)` is a an R package that can read Stata, SAS, and SPSS data into R. This is useful because data providers too often store data in proprietary file formats instead of open formats like .csvs or fixed-width files. Note: while the NBER website does make the .csv version available, please download the Stata file to familiarize yourself with `library(haven)`.

1. Create a new tibble from `data` created in 1.11 (not the survey object) called `cps2021` with just data from 2021.
2. Use `library(haven)` to read `cpsb202104.dta` into R and assign it to `nber`. Drop all variables except `hufinal` and `pwcmpwgt`.
3. Count the number of observations in `nber` and compare it to the number of observations in `cps2021`.
4. Read the documentation for `hufinal` in `cpsb202001.ddf`, which can be found under the documentation page. You may need to open the document in a web browser. Drop all observations with codes for `hufinal` that indicate that the labor force questions were incomplete. (**hint:** You should only need one less than or greater than operator because the order of the codes is meaningful). The number of rows in `cps2021` and `nber` should now match.
5. `cps2021` and `nber` should have a common sort order so combine the columns with `bind_cols()` and assign to `nber_cps`.
6. According to the documentation, `pwcmpwgt` has four implied decimal places. Use `mutate()` to divide `pwcmpwgt` by 10,000.
7. Create `nber_cps_svy` with `as_survey_design()` and use `weight = pwcmpwgt`.
8. Use `mutate()` to create three numeric indicator variables:
   - `labor_force` if the observation is in the labor force.
   - `employed` if the observation is employed.
   - `unemployed` if the observation is unemployed.
9. Use `AGE` and `POPSTAT` to filter to the civilian populations age 16, group by year, and use `survey_total()` to:
   - Calculate the number of people in the labor force with `labor_force`.
   - Calculate the number of people employed with `employed`.
   - Calculate the number of people unemployed with `unemployed`.

10. Compare the results you calculated with microdata to the official tabulation for April 2021. Now your calculations should match the published tables (by the closest 1,000 because of rounding).

## Exercise 05 (2 points)

Create a data visualization with the CPS data using `library(ggplot2)`. Ignore the weights unless you summarize the data before visualization. Make sure to use appropriate visual encodings, and further create a title, subtitle, legend or axis labels, and properly source the data in a caption.