

Data Analysis Project

Exploring the American Landscape with IPUMS & ATUS

dat309

2026-10-19

Table of contents

1	Project Overview	1
2	This Project Ignores Survey Weights	2
3	The Dataset & Core Variables	2
4	Reading and Cleaning the Data	3
5	Interested in downloading your own extract from IPUMS? See the instructions below.	3
6	Potential Research Questions	4

1 Project Overview

In this project, you will step into the shoes of a data analyst. Using individual-level survey microdata from the **IPUMS USA (American Community Survey)**, you will formulate a research question, clean and transform microdata, conduct exploratory data analysis (EDA), and summarize your findings with simple, descriptive statistics and visualizations.

Reproducibility matters in data science. Your final deliverable is a self-contained, reproducible **Quarto document** (`.qmd`) rendered to HTML or PDF.

- **Total Points:** 100 Points
- **Format:** Rendered HTML or PDF report + Quarto source code (`.qmd`)

A note on scope: This project is designed for students who are learning the basics of data wrangling and visualization in R. It intentionally sticks to **unweighted** summary statistics and plots — no survey weighting, no regression modeling. This keeps the focus on `dplyr/ggplot2` fundamentals: filtering, recoding, summarizing, and plotting. See the important caveat below about what this simplification costs you.

2 This Project Ignores Survey Weights

The American Community Survey (ACS) is **not** a simple random sample — some people and places are sampled at higher or lower rates by design. Every respondent carries a person weight, **PERWT**, representing the number of people in the U.S. population that respondent stands in for. To produce statistics that generalize to the actual U.S. population, analysts *must* incorporate **PERWT** (typically with a package like **srvyr**).

In this project, you will ignore **PERWT and compute plain, unweighted statistics** (e.g., `mean(INCTOT)`, not a survey-weighted mean). This is a deliberate simplification to let you focus on core wrangling and plotting skills before tackling survey design.

This means: the numbers you compute in this project (means, proportions, medians, rates) are descriptive summaries of *your extracted sample*, not valid estimates of the U.S. population. Do not report them as if they generalize — e.g., avoid phrasing like “the average U.S. worker earns...”. Instead, phrase your findings as being about *the individuals in this sample* — e.g., “among respondents in this sample...”. You will learn to do this correctly, with **PERWT** and survey design tools, in a later course/project.

3 The Dataset & Core Variables

You will work with a customized extract from the American Community Survey (ACS) distributed via IPUMS USA. Your dataset includes the following primary variables:

Variable Name	Description	Role / Data Type	Key Considerations & Special Codes
INCTOT	Total personal income	Continuous (Numeric)	Top-coded; negative values represent business losses; 9999999 denotes N/A.
DEGFIELD	Field of degree (college major)	Categorical (Unordered)	High cardinality (dozens of codes). Consider collapsing into a handful of broad disciplines (e.g., STEM, Business, Humanities/Arts, Health, Social Sciences).
UHRSWORK	Usual hours worked per week	Discrete / Continuous	00 denotes not working/N/A. Often filtered to full-time workers (≥ 35 hours/week).
TRANWORK	Means of transportation to work	Categorical (Factor)	Includes "Worked from home", public transit, private automobile, walking, bicycle, etc.
TRANTIME	Travel time to work (minutes)	Discrete (Integer)	Valid primarily for individuals who commute (<code>TRANWORK != "Worked from home"</code>). Zero or missing for remote workers.
RENTGRS	Monthly gross rent	Continuous (Numeric)	Only applicable to renters (0 denotes owned or N/A). Combine with INCTOT to calculate a rent-burden ratio.

Variable Name	Description	Role / Data Type	Key Considerations & Special Codes
PERWT	Person survey weight	Numeric Weight	Not used in this project — see the caveat above. Included in the extract so you can see it, and so you're prepared for future coursework that does use it.

Demographic controls available in the extract: AGE, SEX, and RACE.

4 Reading and Cleaning the Data

I recommend starting with one year of data to get your analysis working, then repeating the process for other years. You can combine years later if you want to explore trends over time.

```
library(haven)
library(ipumsr)
library(tidyverse)

# I have pre-saved smaller versions of the data, see below for the full set:

# year 2006
ipums_2006 <- read_rds('https://euclid.nmu.edu/~joshthom/dat309/data/IPUMS/ipums_2006.rds')

# year 2016
ipums_2016 <- read_rds('https://euclid.nmu.edu/~joshthom/dat309/data/IPUMS/ipums_2016.rds')

# year 2024
ipums_2024 <- read_rds('https://euclid.nmu.edu/~joshthom/dat309/data/IPUMS/ipums_2024.rds')

# 3-year combined dataset
ipums_all <- read_rds('https://euclid.nmu.edu/~joshthom/dat309/data/IPUMS/ipums_full.rds')
```

5 Interested in downloading your own extract from IPUMS? See the instructions below.

To load your own extract, you first create it on the IPUMS website, download both the extract's data and the DDI and also set the working directory to the folder with these files (or change the path below).

```
if (!require("ipumsr")) stop("Reading IPUMS data into R requires the ipumsr package. It can be in

# Download & Read the IPUMS extract
```

```
ddi <- read_ipums_ddi("../data/IPUMS/ipums_codebook_full.xml")
ipums <- read_ipums_micro(ddi)
```

From here, all summary statistics and plots should be computed directly on `clean_data` with ordinary `dplyr/ggplot2` functions (`mean()`, `median()`, `count()`, `summarize()`, `geom_bar()`, `geom_histogram()`, etc.) — **do not** use `PERWT` as a weight in any calculation.

6 Potential Research Questions

Select **one** of the following tracks, or formulate an original, question using this variable set. Every track below can be fully addressed with unweighted summary statistics and visualizations — no modeling required.

- **Track A: The College Major and Earnings**

- *Question:* How does undergraduate field of study (`DEGFIELD`) relate to annual personal earnings (`INCTOT`) among respondents in this sample? Do some major groupings show noticeably higher or lower typical earnings than others?

- **Track B: The Remote Work Shift**

- *Question:* Among respondents in this sample, how do demographic and educational factors (`DEGFIELD`, `AGE`, `UHRSWORK`) differ between those who work from home versus those who commute (`TRANWORK`)? For commuters, how does travel time (`TRANTIME`) vary across groups?

- **Track C: Commute Time and Housing Costs**

- *Question:* Among respondents in this sample, do individuals with shorter commute times (`TRANTIME`) tend to pay more in gross rent (`RENTGRS`)? How does rent burden — defined as $(RENTGRS * 12) / INCTOT$ — vary across occupational or educational groups?

- **Track D: Work Hours Across Groups**

- *Question:* How do weekly hours worked (`UHRSWORK`) vary across major disciplines and demographic groups in this sample, and does this line up with differences in typical earnings?

```
install.packages('haven') install.packages('ipumsr')
```

```
ipums
```