Problem Set 1

Commute Times in America

Due: Wednesday, February 11 at 11:59pm

Overview

In this problem set, you'll work with real data from the 2024 American Community Survey (ACS) to explore commute times in the United States. You'll practice setting up a reproducible project, making data cleaning decisions, computing descriptive statistics, and creating visualizations.

Submission: Submit your rendered `ps1.pdf` on Gradescope. Make sure your GitHub URL is included in the document.

---

1 Project Setup (10 points)

1. Create a folder called `gov51-ps1` in your local Projects directory (NOT in Dropbox/iCloud/OneDrive—these cause problems with Git).
2. Create a new RStudio Project in this folder with Git enabled.
3. Create this subfolder structure:

```
gov51-ps1/
  data/
    raw/
  code/
  output/
```

4. **Download data from IPUMS** (see Appendix A for detailed instructions):
   - Go to `https://usa.ipums.org` and create a free account
   - Select the **2024 ACS 1-year sample**
   - Add these variables: `STATEFIP`, `AGE`, `SEX`, `EDUC`, `EMPSTAT`, `TRANTIME`, `INCTOT`
   - Download as CSV, rename to `acs2024.csv`, and place in `data/raw/`
5. Create a new Quarto document called `ps1.qmd` in your project folder (not inside code/).
6. Make your first commit with message: "Initial project setup"
7. Create a GitHub repository and push your project. **Include your GitHub repository URL at the top of your Quarto document** (right after the title/author). We will check your repo to verify folder structure and commits.

---

2 Get to Know Your Data (15 points)

Write code in R chunks and answers in plain text in your Quarto document.

1. Load the tidyverse and read in your ACS data.
2. Using the IPUMS documentation (online at `https://usa.ipums.org`), look up what each variable in your dataset means. Pay special attention to `TRANTIME` (What does a value of 0 mean?) and `PERWT` (What does this represent?).
3. **Create new variables** from the raw IPUMS codes. You'll need to consult the IPUMS documentation to understand what each code means. *Hint*: `mutate()` creates new variables.
   - `female`: equals 1 if female, 0 otherwise (from `SEX`)
   - Education dummies (from `EDUC`): `less_than_hs`, `hs_only`, `some_college`, `college_only`, `advanced_degree`
   - Employment dummies (from `EMPSTAT`): `employed`, `unemployed`, `not_in_labor_force`

   Important: IPUMS uses special codes for missing or not-applicable values (e.g., 0 or 99 may indicate "N/A" rather than a true value). Check the codebook carefully.
4. **Why does this matter?** In 2–3 sentences, explain what could go wrong if you accidentally treated "N/A" codes as real zeros when calculating summary statistics. Give a concrete example using one of the variables in your dataset.
5. **Create a summary statistics table** for your dataset. The table must be generated by R code—do not type the values manually. This is a core principle of reproducible research: if the data changes, your table should update automatically.

   Your table should include these variables:

Table 1: Summary Statistics for 2024 ACS Sample

Variable	N	Mean	Std. Dev.	Min	Max
Age	XXX	XXX	XXX	XXX	XXX
Female	XXX	XXX	XXX	XXX	XXX
Less than High School	XXX	XXX	XXX	XXX	XXX
High School Only	XXX	XXX	XXX	XXX	XXX
Some College	XXX	XXX	XXX	XXX	XXX
College Only	XXX	XXX	XXX	XXX	XXX
Advanced Degree	XXX	XXX	XXX	XXX	XXX
Employed	XXX	XXX	XXX	XXX	XXX
Unemployed	XXX	XXX	XXX	XXX	XXX
Not in Labor Force	XXX	XXX	XXX	XXX	XXX
Commute Time (mins)	XXX	XXX	XXX	XXX	XXX
Total Income ($)	XXX	XXX	XXX	XXX	XXX

Important: Your table should be readable on its own—someone should be able to understand what they're looking at without reading your code. Include a descriptive title, clear variable names (not raw variable codes), and units where appropriate.

The table must render cleanly in your PDF (not raw console output). *Hint*: The `kable()` function from the `knitr` package is the simplest approach. See: https://quarto.org/docs/authoring/tables.html

6. Does the maximum commute time make sense for a daily commute? Write 1–2 sentences.

---

3 Who Should Be in Your Analysis? (25 points)

This is the most important section. You must figure out the logic yourself.

You want to study **commute times for people who actually commute to work**. But the raw data includes everyone—children, retirees, unemployed people, remote workers, etc.

Your task: Decide which observations to keep and which to drop.

1. **Investigate the data.** Answer these questions with code:
   - How many people have `TRANTIME == 0`? What percentage is this?
   - Who are these people with zero commute time? (Think about what categories of people wouldn't have a commute.)
2. **Make your decision.** Create a subset called `commuters` that includes only people who should be in your commute time analysis.

   Write 2–3 sentences explaining your criteria and why.
3. **Verify your subset.**
   - How many observations remain?
   - What is the new minimum TRANTIME? Does this make sense?
4. **Create a new summary statistics table** for your `commuters` subset, using the same format as your table in Section 2. This lets you compare the full sample to your analytical sample.
5. Commit with message: "Create commuters subset"

---

4 Visualize and Interpret (20 points)

Using your `commuters` subset:

1. Create a histogram of commute times using `ggplot2`. Include:
   - An informative title
   - Clear axis labels
   - A reasonable bin width

2. In 2–3 sentences, describe what you see: Is it symmetric or skewed? Where is the center? Are there any notable features (e.g., spikes at certain values)?
3. Using your summary statistics table from Section 3, compare the mean and median commute time. Which is larger? What does this tell you about the shape of the distribution, and does it match what you see in the histogram?
4. Save your histogram to `output/commute_histogram.png` using `ggsave()`.
5. Commit with message: "Add histogram"

---

5 The Weight of Evidence (15 points)

Survey data comes with **weights**. The variable `PERWT` tells you how many people in the U.S. each respondent represents.

- **Without weights**: mean = average among people *in the sample*
- **With weights**: mean = average among people *in the United States*

1. Calculate the **weighted mean**:

```
weighted.mean(commuters$TRANTIME, commuters$PERWT)
```

2. Compare: Is the weighted mean different from the unweighted mean you calculated earlier? By how much?
3. In 2–3 sentences, explain what each number represents and why they might differ.

Submission Checklist

Before submitting:

☐ RStudio Project with correct folder structure

☐ `ps1.qmd` renders to PDF without errors

☐ Summary statistics table for full ACS sample

☐ `commuters` subset with written justification

☐ Summary statistics table for commuters

☐ Histogram saved to output/

☐ Weighted vs. unweighted comparison

☐ At least 3 commits with meaningful messages

☐ Pushed to GitHub

Submit ps1.pdf on Gradescope (with your GitHub URL visible in the document).

Grading

Section	Points
1. Project Setup	10
2. Get to Know Your Data	20
3. Who Should Be in Your Analysis?	25
4. Visualize and Interpret	20
5. The Weight of Evidence	25
Total	**100**

A IPUMS Data Download Instructions

IPUMS USA (https://usa.ipums.org) provides free access to U.S. Census microdata. Follow these steps to download the data for this assignment.

Step 1: Create an Account

1. Go to https://usa.ipums.org/usa/
2. Click "Login" → "Create an account"
3. Use your Harvard email, select "Coursework" as intended use
4. Verify your email (usually instant)

Step 2: Start a Data Extract

1. Click the green "Get Data" button
2. Click "SELECT SAMPLES"
3. Uncheck "Default sample from each year"
4. Check **only** "2024 ACS"
5. Click "SUBMIT SAMPLE SELECTIONS"

Step 3: Select Variables

Click "SELECT HARMONIZED VARIABLES" and add these variables to your cart:

Category	Variable	Description
HOUSEHOLD → GEOGRAPHIC	`STATEFIP`	State code
PERSON → DEMOGRAPHIC	`AGE`	Age
PERSON → DEMOGRAPHIC	`SEX`	Sex
PERSON → EDUCATION	`EDUC`	Educational attainment
PERSON → WORK	`EMPSTAT`	Employment status
PERSON → WORK	`TRANTIME`	Commute time (minutes)
PERSON → INCOME	`INCTOT`	Total personal income

To add each variable: navigate to its category, click the variable name, click "ADD TO CART."

Step 4: Create and Download the Extract

1. Click "VIEW CART" → "CREATE DATA EXTRACT"
2. Set description to "Gov 51 PS1"
3. **Important**: Change data format to "Comma delimited (.csv)"
4. Click "SUBMIT EXTRACT"
5. Wait 1–5 minutes, then download the `.csv.gz` file

6. Unzip it (double-click on Mac, use 7-Zip on Windows)
7. Rename to `acs2024.csv` and move to your `data/raw/` folder

Troubleshooting

- **Extract taking too long?** Large extracts can take 10–30 minutes. You'll get an email when ready.
- **Can't find a variable?** Use the search box on IPUMS.
- **Account not approved?** Check spam folder. Usually instant but can take up to 24 hours.