Problem Set 1 — Amendment

Handling Large Data Files with Git

February 6, 2026

Important Notice

If you are having trouble pushing your project to GitHub because of the `acs2024.csv` file, this document explains the problem and how to fix it.

The Problem

GitHub has a **100 MB file size limit**. The ACS 2024 data file you downloaded from IPUMS is larger than this limit, so Git will refuse to push it to your remote repository.

You may see an error message like:

```
remote: error: File data/raw/acs2024.csv is 150.00 MB;
this exceeds GitHub's file size limit of 100.00 MB
```

Why This Happens (And Why It's Actually Normal)

This is not a mistake—it's a **common situation in real research**. Data files are often too large to store in version control systems like Git. Professional researchers typically:

- Store **code** in Git (version controlled, shared on GitHub)
- Store **large data files** separately (locally, or on cloud storage like Dropbox/Google Drive)
- Document where collaborators can obtain the data

You're learning a real-world workflow!

The Solution

We will tell Git to **ignore** the large data file. Your code, output, and folder structure will be pushed to GitHub, but the data file will stay only on your local computer.

Step 1: Create a `.gitignore` file

In the **root of your project folder** (the same level as your `.Rproj` file), create a new text file called `.gitignore` (note the dot at the beginning).

Add this line to the file:

```
# Large data files - keep locally, don't push to GitHub
data/raw/acs2024.csv
```

Save the file.

Step 2: If you already tried to commit the data file

If you already added or committed the large file, you need to remove it from Git's tracking (this does NOT delete the file from your computer):

Open the Terminal tab in RStudio (next to the Console tab) and run:

```
git rm --cached data/raw/acs2024.csv
```

Step 3: Commit and push

Now commit your `.gitignore` file and push everything:

```
git add .gitignore
git add -A
git commit -m "Add gitignore for large data file"
git push
```

Or use the RStudio Git pane:

1. Check the box next to `.gitignore`
2. Check boxes for any other files you want to commit
3. Click "Commit", write your message, click "Commit"
4. Click "Push"

What Your Repository Should Look Like

After pushing, your GitHub repository should contain:

```
gov51-ps1/
  .gitignore           # Tells Git what to ignore
  gov51-ps1.Rproj      # Your R project file
  ps1.qmd              # Your Quarto document
  data/
    raw/               # This folder exists but is empty on GitHub
  code/                # Any R scripts
  output/              # Your figures, tables, etc.
```

The `data/raw/` folder will appear empty on GitHub, but on your local computer it still contains `acs2024.csv`. That's correct!

Will My Code Still Work?

Yes! Your R code uses **relative paths** like:

```
acs <- read_csv("data/raw/acs2024.csv")
```

This path works on your local computer because the file is still there. The only difference is that Git doesn't track or push the file.

Quick Reference

Your `.gitignore` file should contain:

```
# Large data files
data/raw/acs2024.csv
```

If you already committed the file, run:

```
git rm --cached data/raw/acs2024.csv
```

Then commit and push as normal.

Questions? Post on Slack, email Harrison, George, or Scott, or come to office hours.