DATA SCIENCE FOR ECONOMISTS

ECON 220 LAB

Jafet Baca-Obando

Week 9, Handling IPUMS Data (Part 2) – 10/24/2025

Outline

01 IPUMS

02 Recoding and outlier analysis

03 Basic sampling

What is IPUMS?

Integrated Public Use Microdata Series

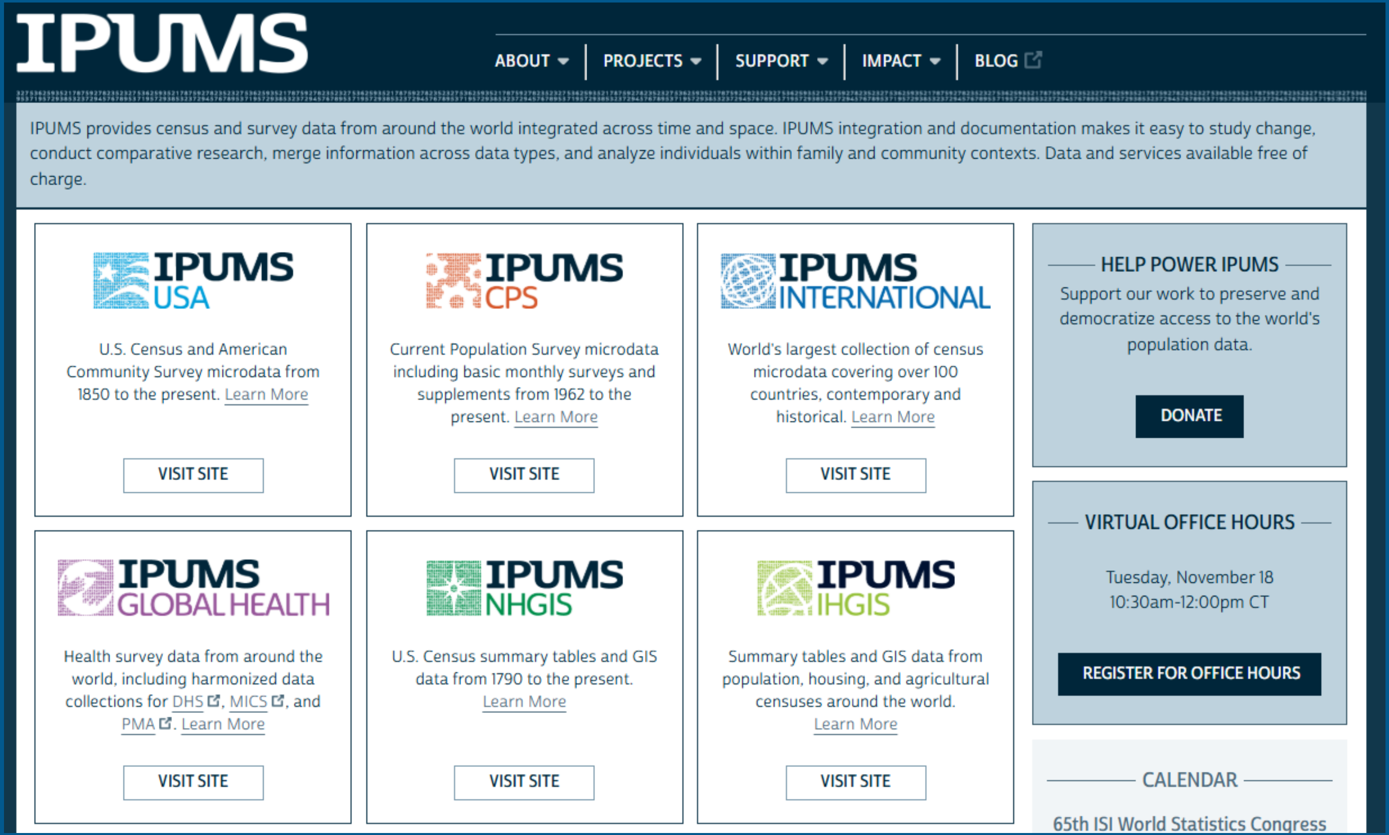
Operated by the University of Minnesota

Core mission: data harmonization

IPUMS takes datasets that were originally collected with different questions, codes, and variable names and makes them consistent. Lots of recoding!

Free access!

By providing access to detailed, anonymized individual-level data (microdata), IPUMS allows researchers to ask complex questions that can't be answered with aggregated summary tables.



IPUMS provides census and survey data from around the world integrated across time and space. IPUMS integration and documentation makes it easy to study change, conduct comparative research, merge information across data types, and analyze individuals within family and community contexts. Data and services available free of charge.

IPUMS USA

U.S. Census and American Community Survey microdata from 1850 to the present. Learn More

VISIT SITE

IPUMS CPS

Current Population Survey microdata including basic monthly surveys and supplements from 1962 to the present. Learn More

VISIT SITE

IPUMS INTERNATIONAL

World's largest collection of census microdata covering over 100 countries, contemporary and historical. Learn More

VISIT SITE

HELP POWER IPUMS

Support our work to preserve and democratize access to the world's population data.

DONATE

IPUMS GLOBAL HEALTH

Health survey data from around the world, including harmonized data collections for DHS, MICS, and PMA. Learn More

VISIT SITE

IPUMS NHGIS

U.S. Census summary tables and GIS data from 1790 to the present. Learn More

VISIT SITE

IPUMS IHGIS

Summary tables and GIS data from population, housing, and agricultural censuses around the world. Learn More

VISIT SITE

VIRTUAL OFFICE HOURS

Tuesday, November 18
10:30am-12:00pm CT

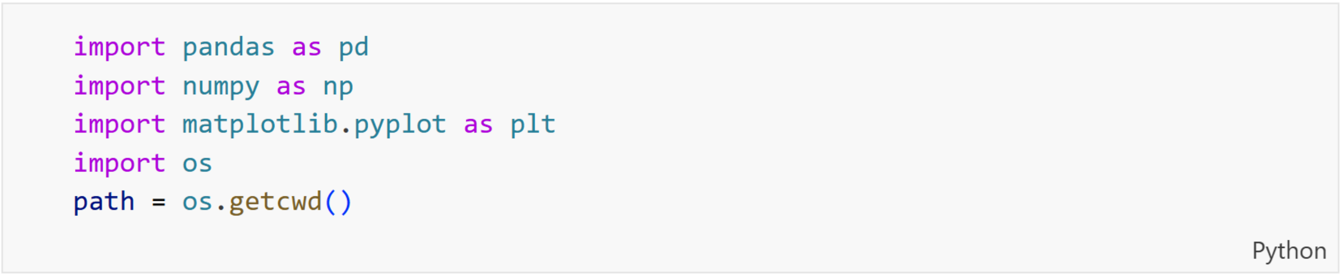
REGISTER FOR OFFICE HOURS

CALENDAR

65th ISI World Statistics Congress

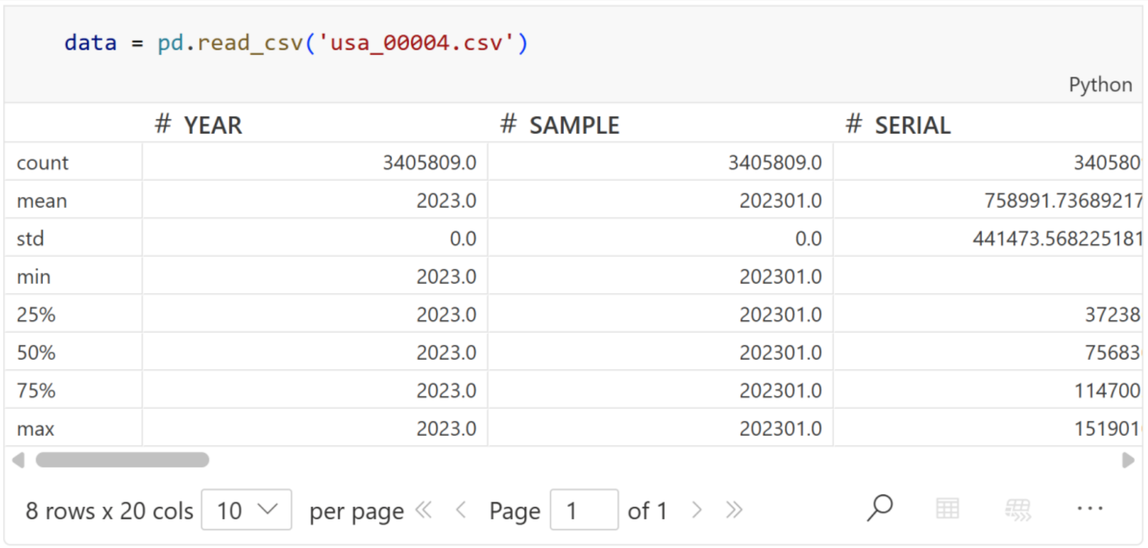
- How to register and extract data? Check guide on Canvas.

Importing required libraries



```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import os
path = os.getcwd()
```

Load and explore the data



```python
data = pd.read_csv('usa_00004.csv')
```

Python

	# YEAR	# SAMPLE	# SERIAL
count	3405809.0	3405809.0	340580
mean	2023.0	202301.0	758991.73689217
std	0.0	0.0	441473.568225181
min	2023.0	202301.0	
25%	2023.0	202301.0	37238
50%	2023.0	202301.0	75683
75%	2023.0	202301.0	114700
max	2023.0	202301.0	151901

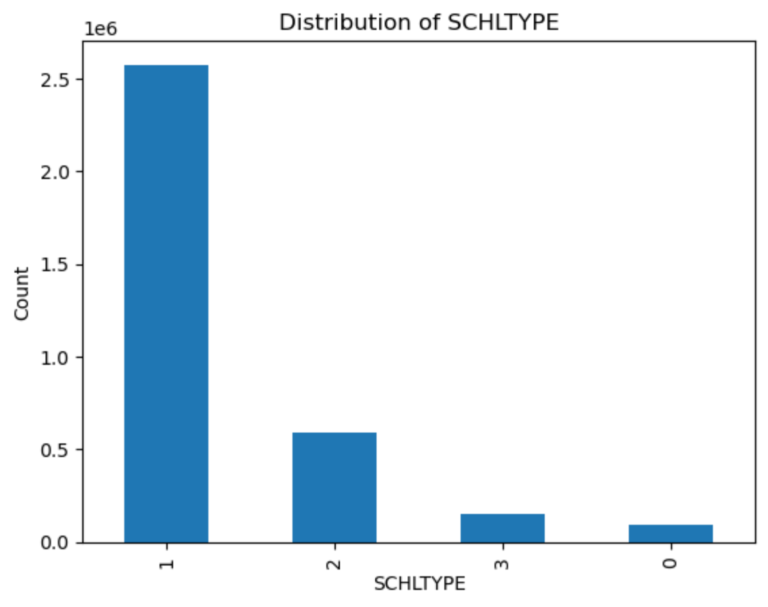
8 rows x 20 cols 10 per page « ‹ Page 1 of 1 › » ···

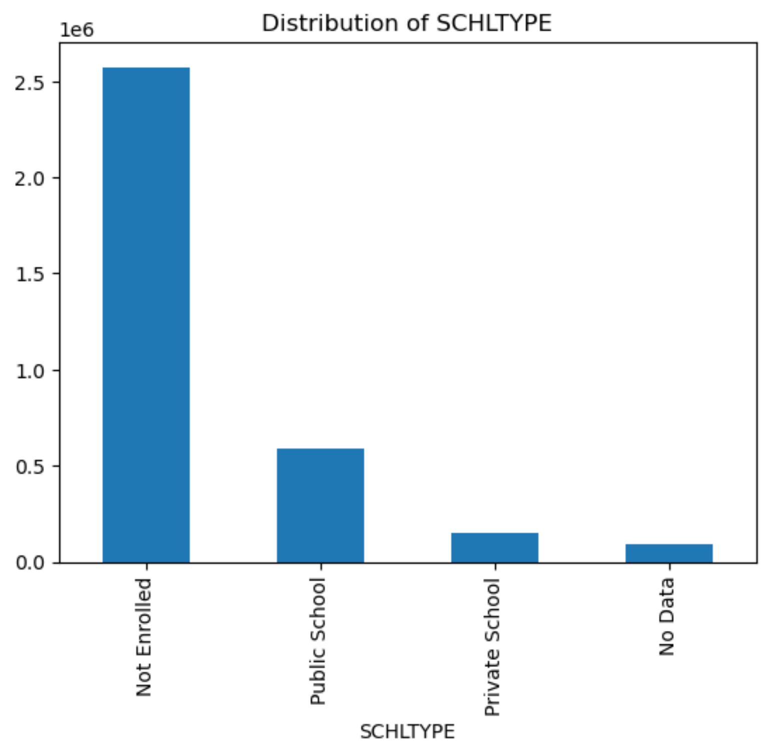
- Note the large number of observations!

Recoding a categorical variable

- IPUMS uses numeric codes for categorical variables.
- **Example:** SCHLTYPE (School Type).
- **Need the Codebook!** Tells us: 0=N/A, 1=Not Enrolled, 2=Public, 3=Private.
- **Recoding**: replace numbers with meaningful labels.

Recoding a categorical variable





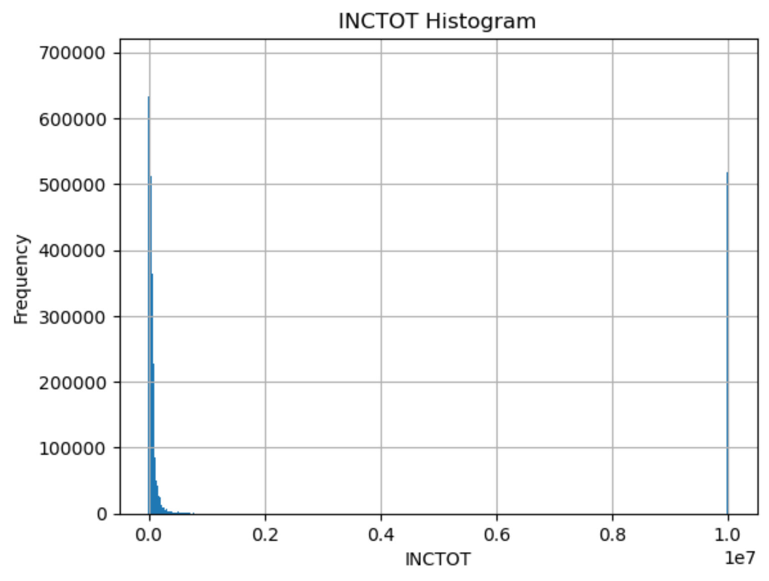
Outlier detection and treatment

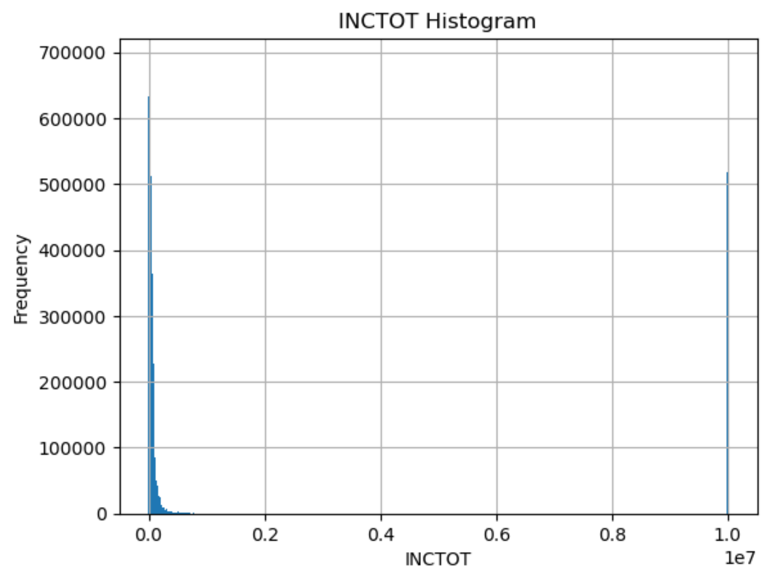
- **Outliers:** Extreme values that might distort analysis.
- Idea:
  - Use a histogram (.hist()) to see the distribution.
  - Observe asymmetry and potential extreme high/low values (including negative).
- Treatment:
  - Interquartile range (IQR) =Q3 (75th percentile) - Q1 (25th percentile)
  - **Define Bounds:** Lower = Q1 - 1.5*IQR, Upper = Q3 + 1.5*IQR.
  - **Filter:** Keep data only within these bounds.

Outlier detection and treatment

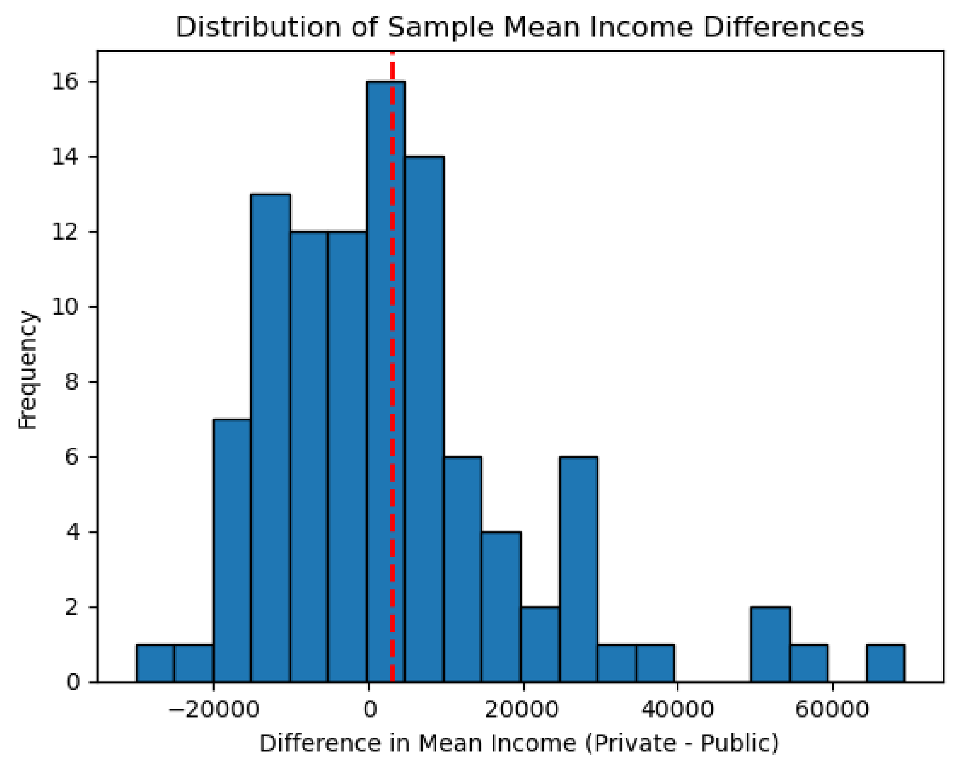
- IQR might not catch all problematic values (e.g., 0 or negative income).
- **Domain knowledge**: Is zero/negative income plausible/useful for this analysis?
- **Filter**: Remove observations with INCTOT <= 0.

Outlier detection and treatment





Random sample



Recap

- **Check the Codebook**: It's essential for understanding IPUMS numeric codes.
- **Recode Variables**: Use .replace() to make your data readable (e.g., 2 → "Public School").
- **Find Outliers**: Use .hist() to visualize data and spot extreme values.
- **Sample for Speed**: Use .sample() for quick analysis on large datasets, but be aware of sampling variability.

To-do list

- **Complete Data Exercise 6**
    - Upload Jupyter notebook (.ipynb file) and HTML file on **October 26**
- **Complete Data Exercise 7**
    - Upload Jupyter notebook (.ipynb file) and HTML file on **November 2**