Analysis of census data

James, Carmen, Jinwoo

Agenda

- Central Question
- Object of Study
- 3 Approaches
- Conclusion

Object of Study : What is this dataset?

- 48,842 records
- 14 demographic and employment features
- Binary income classification (>50K or <= 50 K)
- Derived from 1994 U.S. Census
- Designed for predictive classification

Problem Statement

The Census income dataset simplifies income into two categories and reduces complex social identities into fixed demographic variables. How do their structural choices influence what form of inequality of inequality can be represented, and what important indicators of economic and social identity are excluded from the dataset.

Method 1: Koopman Format Anatomies

Format Anatomies - Meso Level

While a macro-level analysis would examine the broader institutional and governmental context of the U.S. Census system, our project focuses primarily on the meso and micro levels, where inequality is structurally formatted within the dataset itself.

Koopman's format anatomies method asks us to analyze how identity is constructed through the structure of the dataset. Individuals are forced into predefined demographic categories. Race and sex are limited to fixed options. Education is grouped into standardized levels.

The dataset determines what kinds of identity are administratively recognizable. Anything that falls outside those categories cannot be expressed within the dataset structure.

Format Anatomies - Micro level (Binary Formatting)

We examine specific formatting decisions. The most significant example is the binary income classification. Instead of capturing gradations of income, the dataset draws a sharp dividing line at $50,000.

This formatting encourages predictive modeling rather than explanatory analysis. It frames inequality as a threshold-based classification task rather than a continuum shaped by systemic forces.

Method 2: Poirier
Reading Datasets

3 Approaches

Denotative Reading (What are the formal definitions of the variables?):

Unit of observation: Adult individuals recorded in the 1994 U.S. Census database.
Primary variables: age, workclass, education level, marital status, occupation, race, sex, hours per week, native country, etc.
Variable definitions: Come from the U.S. Census Bureau. They include standardized labels for race, binary sex classifications, binary income labels (<50k<)

Connotative Reading (How have definitions changed over time?):

Outdated census categories:

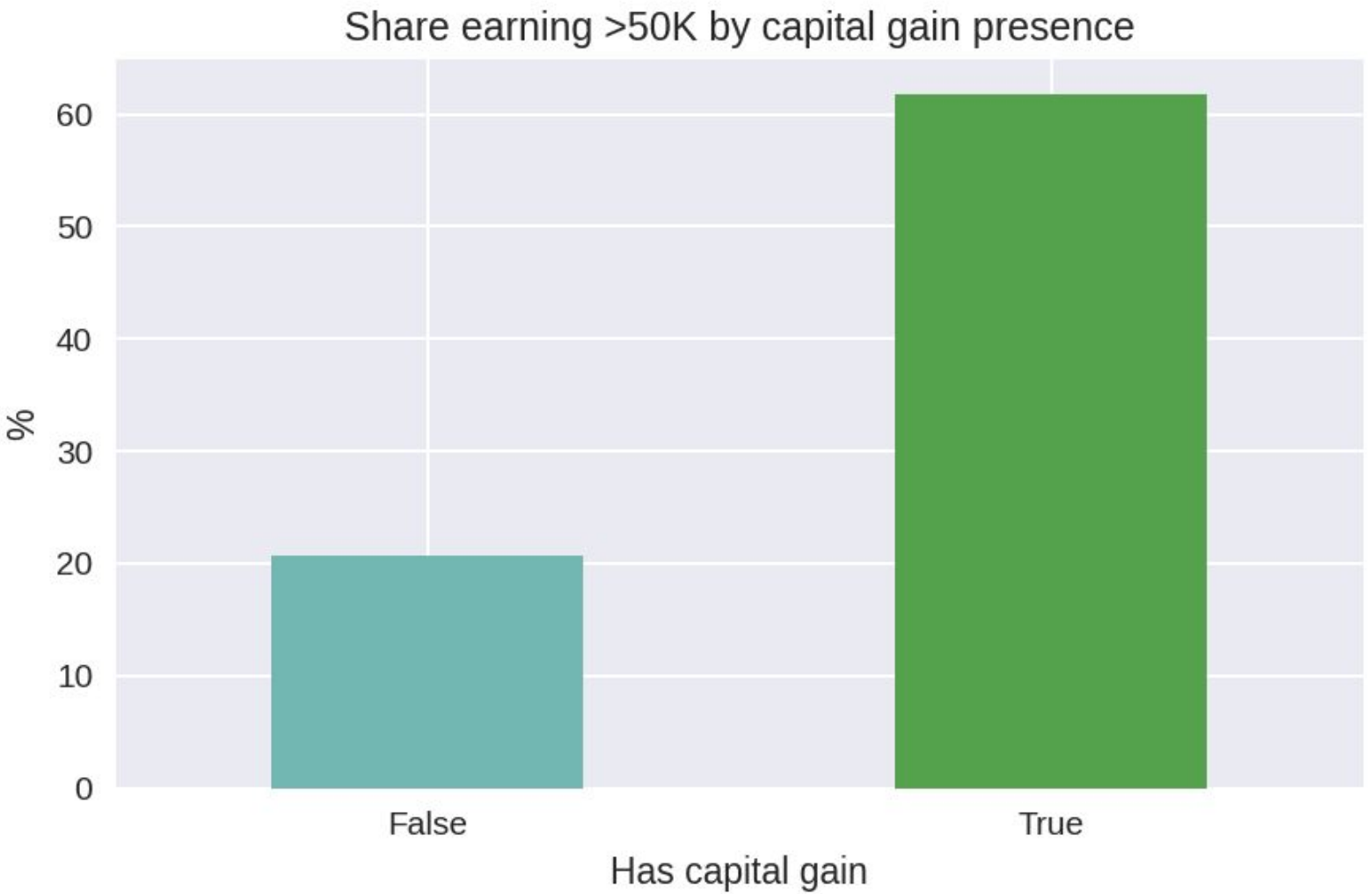
- Race labels (only one option, outdated/missing terms)
- Binary sex classification
- No room for modern forms of employment (Gig work, Hybrid/Remote work)

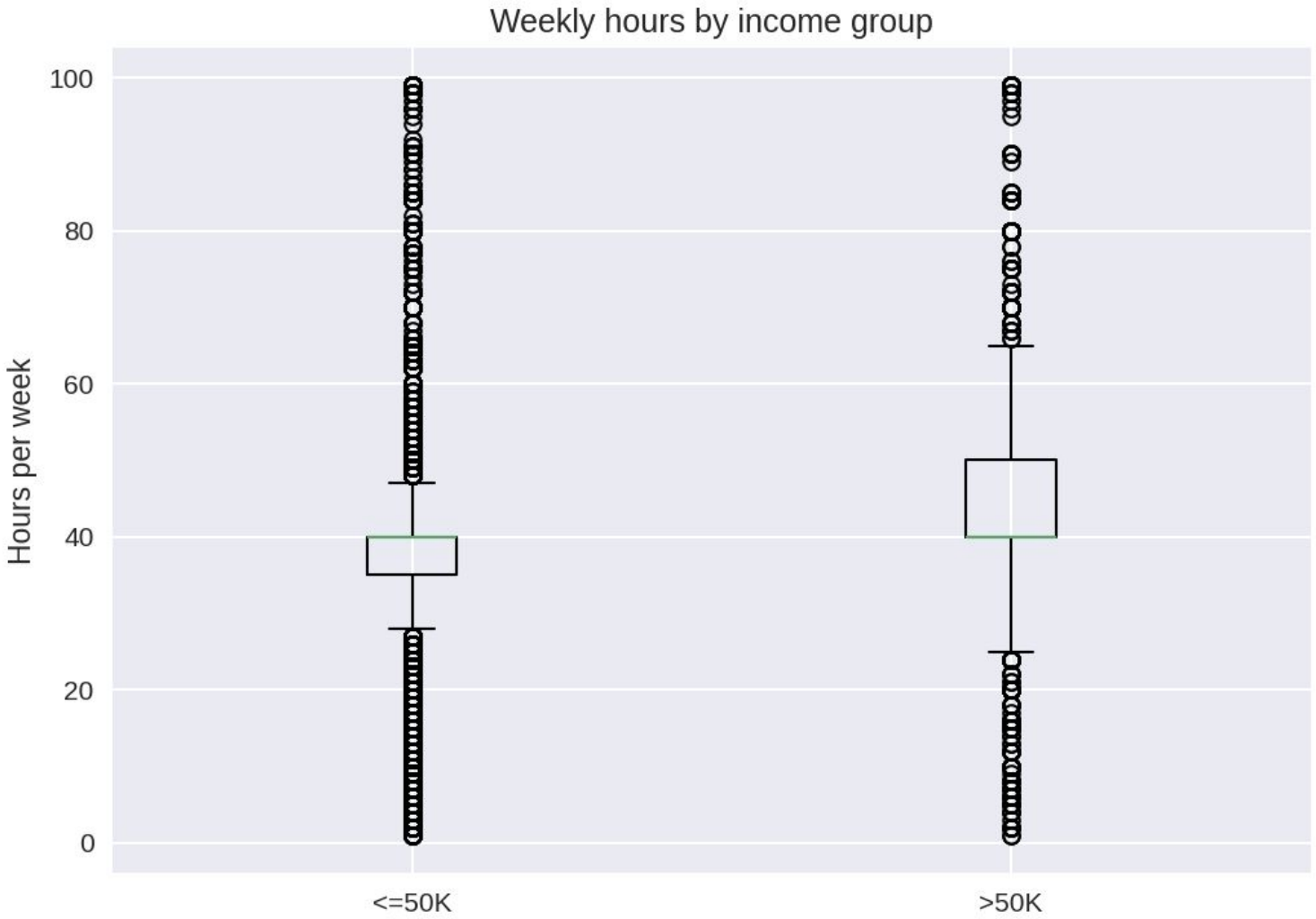
Machine Learning Applications:

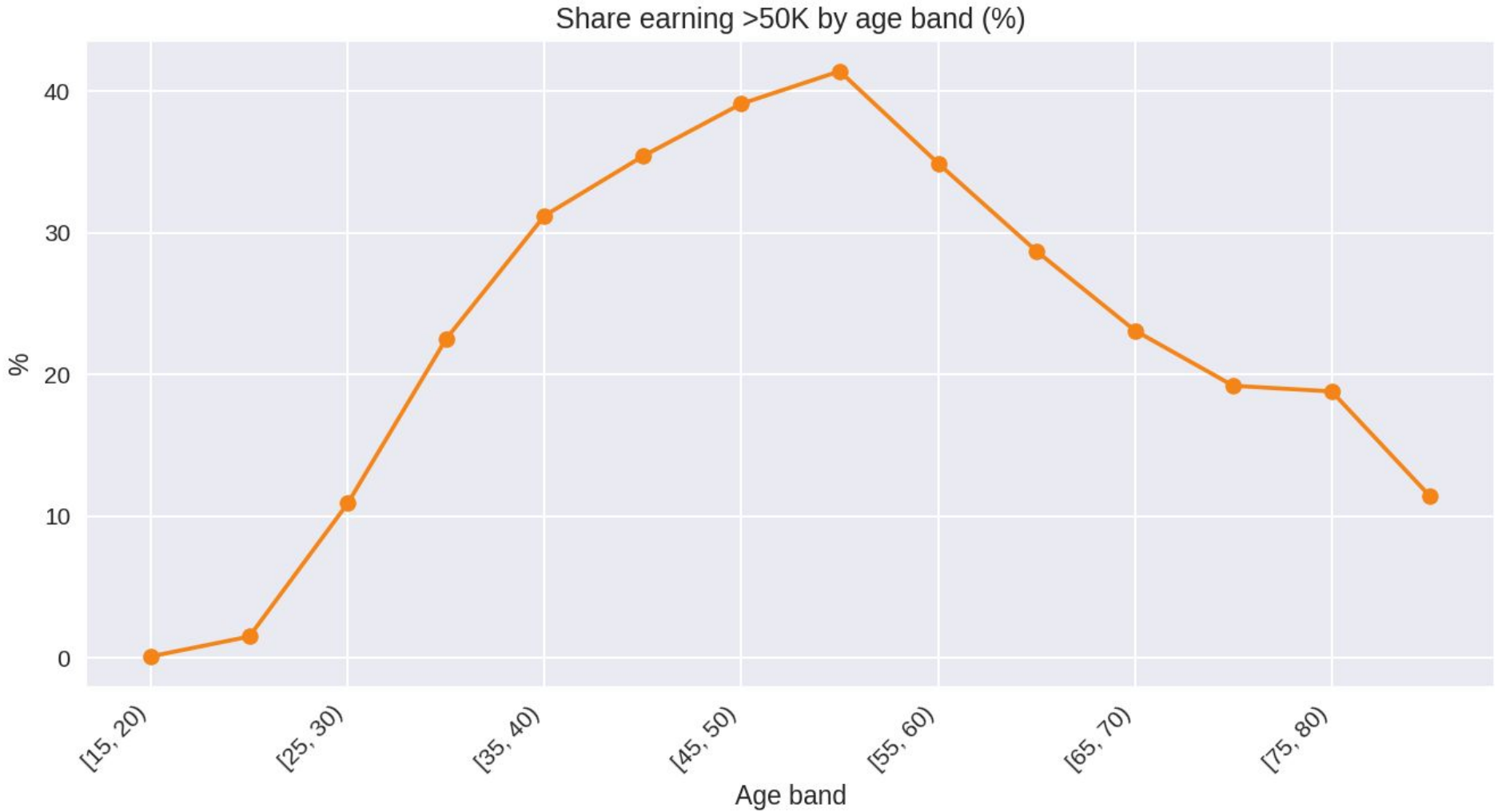
- Simplified 50K income cutoff
- Rigid categorical structures.

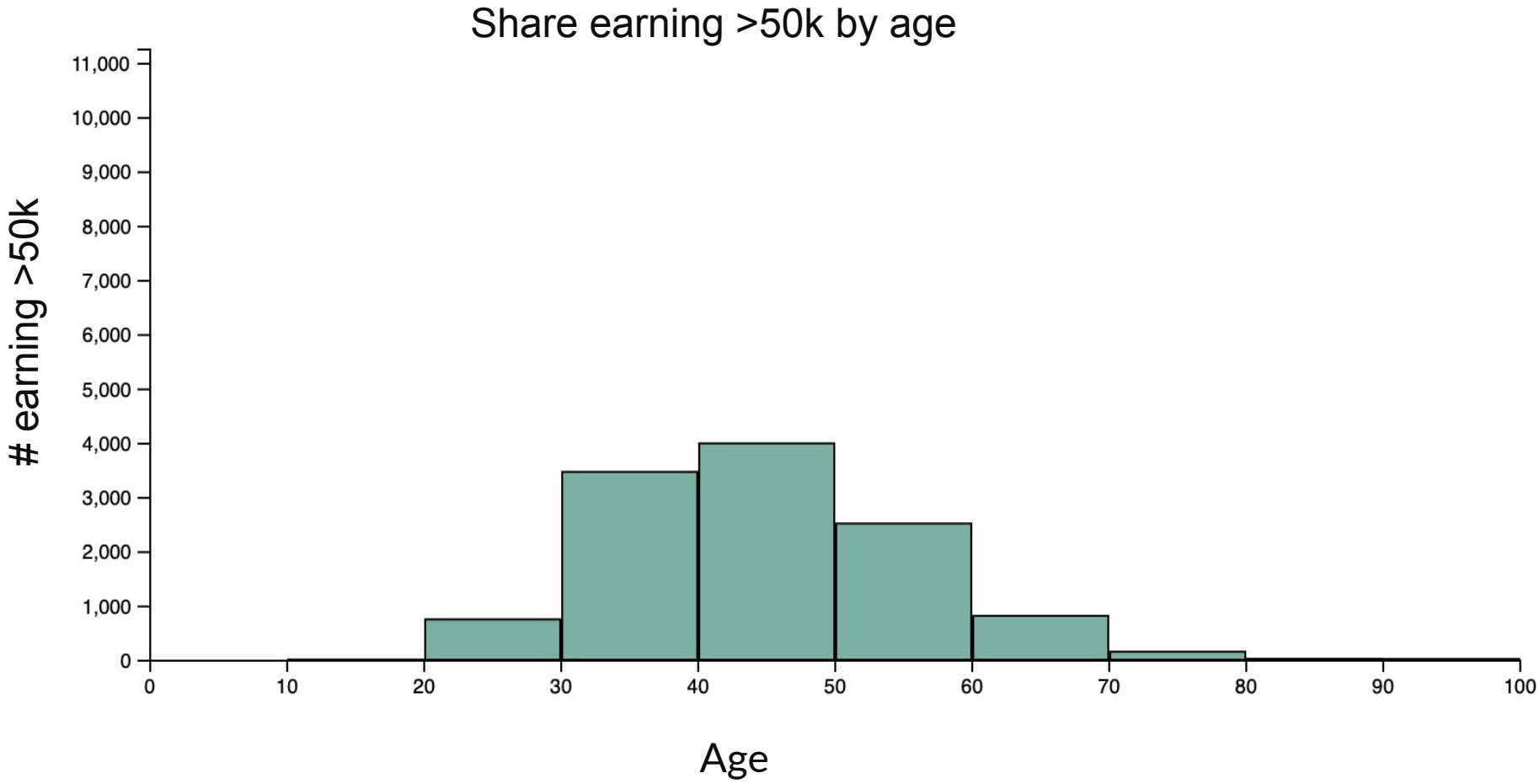
Deconstructive Reading (What is missing? What does this mean for the data?)

Income inequality: Demographic inequality vs. structural inequality
Missing: There are vague categorical variables, missing many aspects of a person's life.


Share earning >50K by capital gain presence
60
50
40
30
20
10
0
%
False
True
Has capital gain


Weekly hours by income group
Hours per week
100
80
60
40
20
0
<=50K
>50K


Share earning >50K by age band (%)
40
30
20
10
0
%
[15, 20)
[25, 30)
[35, 40)
[45, 50)
[55, 60)
[65, 70)
[75, 80)
Age band


Share earning >50k by age
earning >50k
11,000
10,000
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
0
10
20
30
40
50
60
70
80
90
100
Age

Conclusion

The Census Income dataset appears neutral and objective, but its structural and formatting choices shape how inequality can be represented. By reducing income to a binary threshold and identity to fixed categories, the dataset excludes important dimensions of economic and social complexity.

Through Poirier and Koopman's methods, we show data does not simply reflect reality.Therefore, critical examination of datasets is essential for democratic understanding.