

# UCI Adult Income Dataset Analysis

---

Assessing Data Quality for 2025 Business Analytics

**Focus:**

**Timeliness & Completeness**

Sachin M

# Dataset overview

---

Source: 1994 Census Bureau Database (Ronny Kohavi & Barry Becker)

- 15 columns: workclass, education, occupation, sex, income
- Used for predicting income levels ( $\leq \$50K$  or  $> \$50K$ )
- Potential applications in business analytics and demographic analysis

Question: Can 30-year old data still be viable for modern business decisions

# Why Data Quality Matters

---

*"Garbage in, garbage out."*

Consequences of poor data quality:

- Unreliable or biased model predictions (Population growth rates)
- Misleading conclusions and lost revenue (Employee record)
- Missed opportunities and unfair decisions (Telephone survey)

# Dimension 1: Timeliness

— *How promptly data is recorded and remains relevant*

Critical issues:

- Inflation
  - \$1 in 1994 = \$2.17 in 2025
  - \$50K threshold balloons to \$108,500 when adjusted (Federal Reserve Bank of Minneapolis, 2025)
- Women's education
  - Participation increased from <25% to 47% (Hurst, K. 2024)
  - Sampling bias from Current Population Survey?

# Dimension 2: Completeness

— *Ensuring required data is present and sufficient*

- Binary income brackets:
  - ≤\$50K or >\$50K only
  - \$60k vs \$1M
- Missing context
  - 42% of borrowers with at least \$25,000 of loan debt (Federal Reserve, 2024)
  - Unknown values

# Real-world impact of data quality

---

- Salary calculations
  - Market salaries
  - Forgotten employees
- Education trends among the population
  - Women's education

# Recommendations

---

1. Retire the dataset for current-year prediction models
2. Adjust for inflation if historical analysis is necessary
3. Treat unknown values explicitly
  - a. Statistical approaches to identify data collection errors

# Conclusion

---

The 1994 UCI Adult Income Dataset is NOT  
suitable for 2025 business use

# Or is it?

---

- Trading systems
  - ELT vs ETL
- Bias is sometimes needed

# Final thoughts

---

- One size does not fit all
- Problems are puzzles
- Skilled data scientists are good at solving those puzzles

---

—

Thank you!