1. Introduction

- Adult Income dataset of UCI Machine Learning Repository has approximately 48,842 records on demographics and an income variable (<=50K or >50K).
- This project measures data quality dimensions (Completeness, Validity, Bias/Fairness) in order to point out some problems which, influence downstream data mining activities.
- In the analysis, it relates to the concepts of Week 10 (Data Warehousing) and Week 11 (Data Mining) in terms of reproducibility and transparency.
- The scripts and results have been placed in a GitHub repository that can be found here: https://github.com/navdeepnathowal-coder/COMP331-Final-Project.git

2. Data Quality Analysis

Completeness

- The missing values were found in work class, occupation, and native-country.

- Evidence: the percentages of missing entries are indicated in missingsummary.csv.

- Interpretation: These gaps can reflect unfinished ETL processes or the survey non-responses which minimizes the reliability of the datasets.

Validity

- Age and hours per week were range checks, and income and sex were category checks, which showed no invalid values.

- Evidence: invalidvaluecounts.csv had one hundred and zero invalid entries.

Uniqueness

- Checking duplicates there were no exact duplicates.
- Evidence: duplicaterowssamples.csv contained nothing.
- Interpretation: Records are unique, meaning that a row level integrity is reliable.

Bias & Fairness

- There is unequal distribution of income between sex and race lines.

Evidence:

- incomebysex.png indicates that greater percentage of over 50 K income is higher among men.

- incomeby_race.png demonstrates that there are differences in racial categories.

3. Recommendations

- Treatment of missing values pick up the median /mode, or domain-specific imputations.

- Balance dataset: use resampling methods (SMOTE, under sampling) to decrease imbalance of classes.

- Audit fairness: Compare the models between subgroups to make sure they perform equally.

- Automate ETL validations: Insert rules of validation of ranges, categories and missing values.

- Prefer bias reduction: On the one hand, fairness problems should be resolved initially, since they directly influence ethical performance.

4. Conclusion

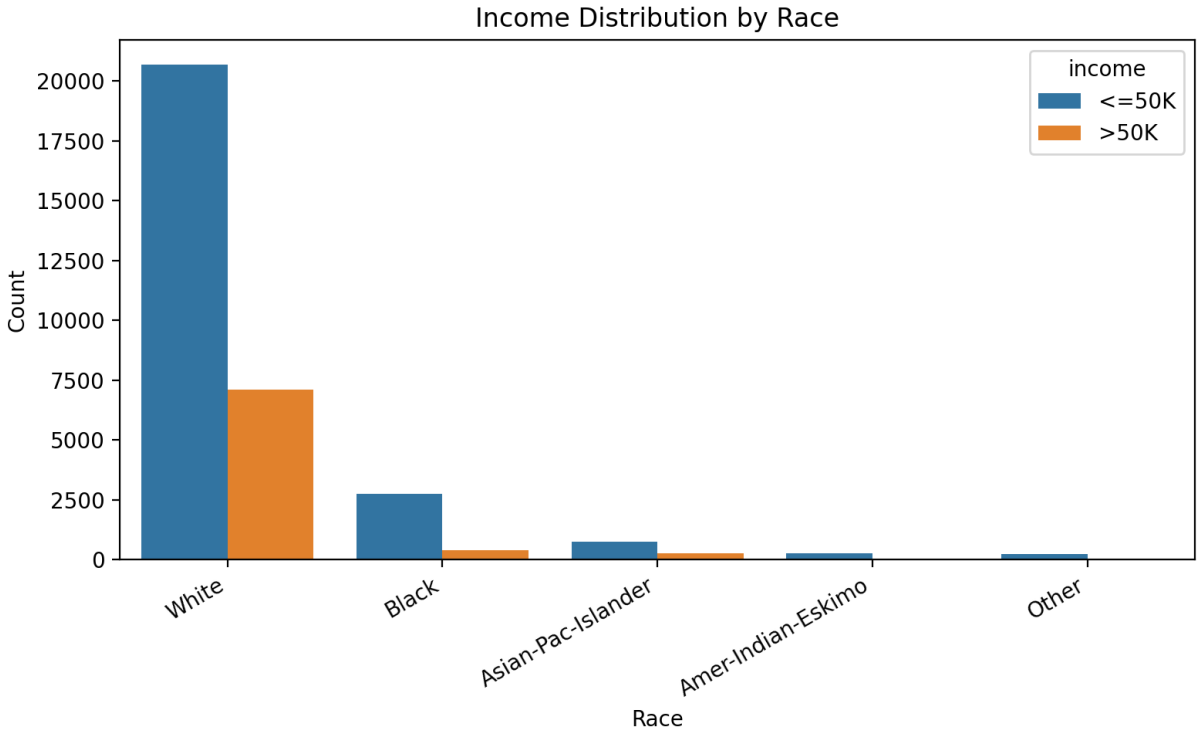
- The Adult Income dataset has high validity and distinctiveness however with missing values and bias/fairness problem.

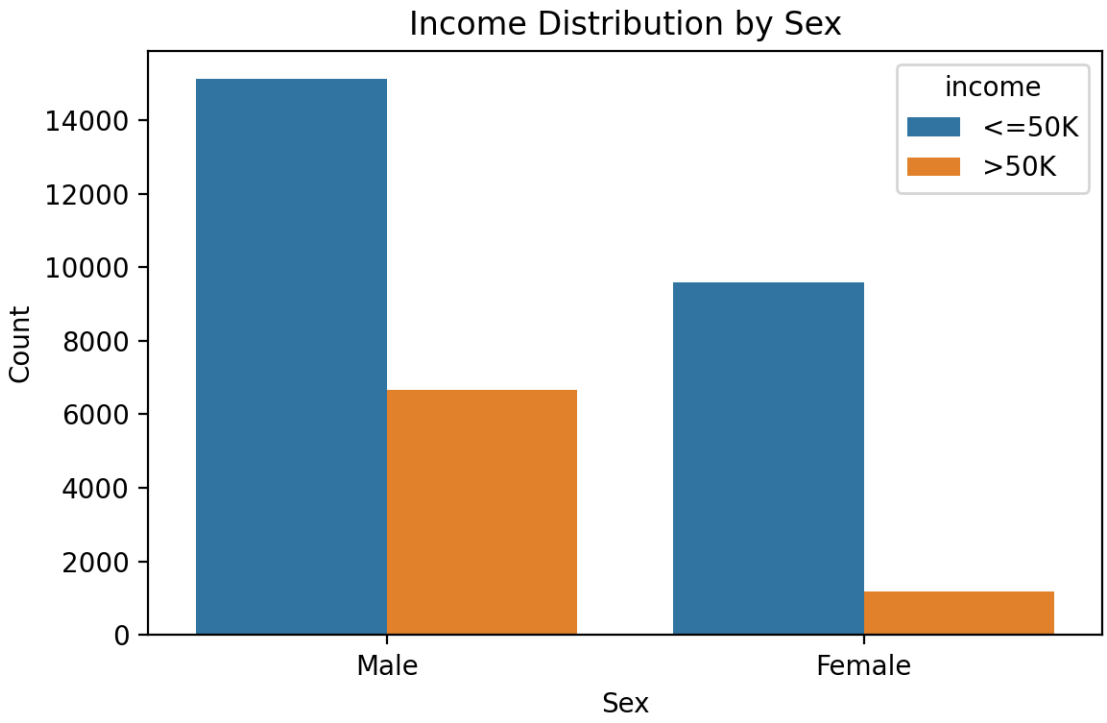
- Lesson learned: It is possible to have structural biases even in clean datasets, which can influence further analysis.

- This project illustrates the significance of data quality profiling to data mining in order to achieve credible and justified results.

References

1. UCI Machine Learning Repository – Adult Dataset.
2. COMP331 Lecture Notes (Week 10: Data Warehousing, Week 11: Data Mining).
3. Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*.
4. Relevant academic papers on bias and fairness in ML.




Income Distribution by Sex
income
<=50K
>50K
Count
14000
12000
10000
8000
6000
4000
2000
0
Male
Female
Sex