Course: Information Extraction, Retrieval and Integration

Unit 4: Data Integration

Assignment Description: Data Integration, Bias and Fairness

Mari Carmen Suárez de Figueroa Baonza
March 2023
Universidad Politécnica de Madrid

General Issues

- This assignment is not the typical one.
  - We can say that it is a **testing-exploratory assignment**
- The objective of this assignment is **to identify the most important difficulties when integrating data**, taking into account **bias and fairness dimensions**.
  - Only a subsets of steps in the data integration process are going to be performed in this assignment
- **Final result** of the assignment is not an integrated dataset, but a set of lessons learned, difficulties, suggestions for improving the process, ways to proceed, among other experiences and knowledge gained.

Assignment Phases

1. Search for and select datasets

- **2 or more datasets** that contain overlapping data
- **Note**: Selected datasets should have some data related to attributes that can caused bias (such as gender, age, race, among others)

2. Identify conflicts among selected datasets

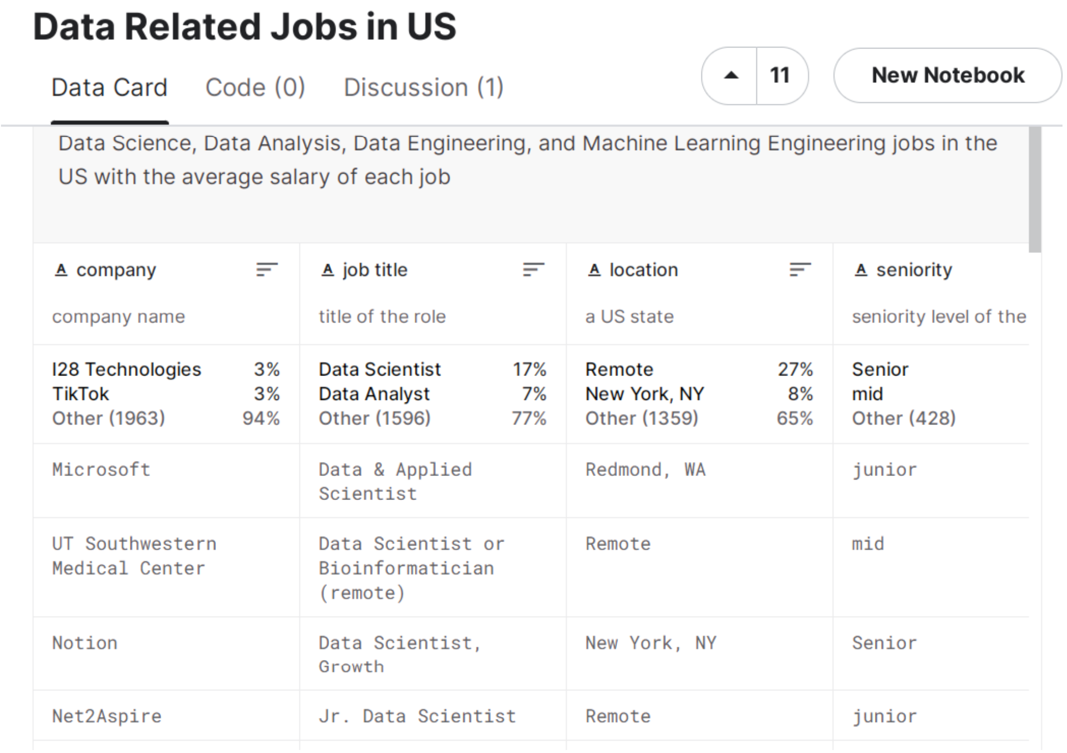
- **Data-level and schema-level** conflicts

3. Be aware about bias and fairness in your datasets

- Use any **tool for mitigating bias**
- Use any **tool for identifying fairness**

1. Search for and select datasets: Example

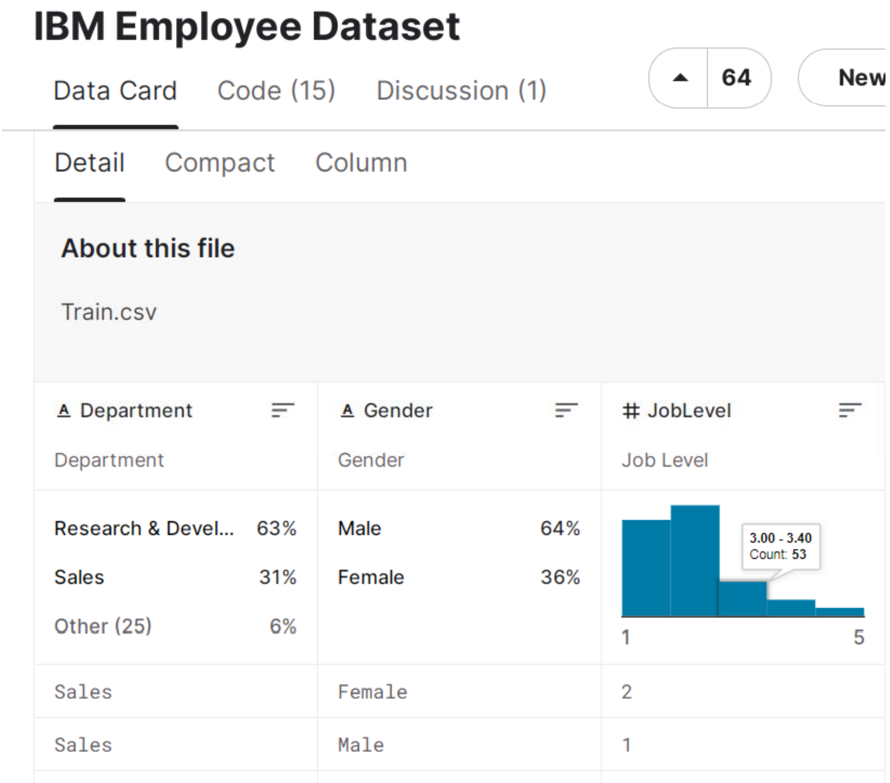
- **Dataset 1**: Data Related Jobs in US
  - https://www.kaggle.com/datasets/mohamedsiika/data-related-jobs-in-us
- **Dataset 2**: IBM Employee Dataset
  - https://www.kaggle.com/datasets/rohitsahoo/employee



Data Related Jobs in US

Data Science, Data Analysis, Data Engineering, and Machine Learning Engineering jobs in the US with the average salary of each job

company	job title	location	seniority
company name	title of the role	a US state	seniority level of the
I28 Technologies 3% TikTok 3% Other (1963) 94%	Data Scientist 17% Data Analyst 7% Other (1596) 77%	Remote 27% New York, NY 8% Other (1359) 65%	Senior mid Other (428)
Microsoft	Data & Applied Scientist	Redmond, WA	junior
UT Southwestern Medical Center	Data Scientist or Bioinformatician (remote)	Remote	mid
Notion	Data Scientist, Growth	New York, NY	Senior
Net2Aspire	Jr. Data Scientist	Remote	junior



IBM Employee Dataset

About this file

Train.csv

Department	Gender	# JobLevel
Department	Gender	Job Level
Research & Devel... 63% Sales 31% Other (25) 6%	Male 64% Female 36%	
Sales	Female	2
Sales	Male	1

1. Search for and select datasets: Dataset Repositories

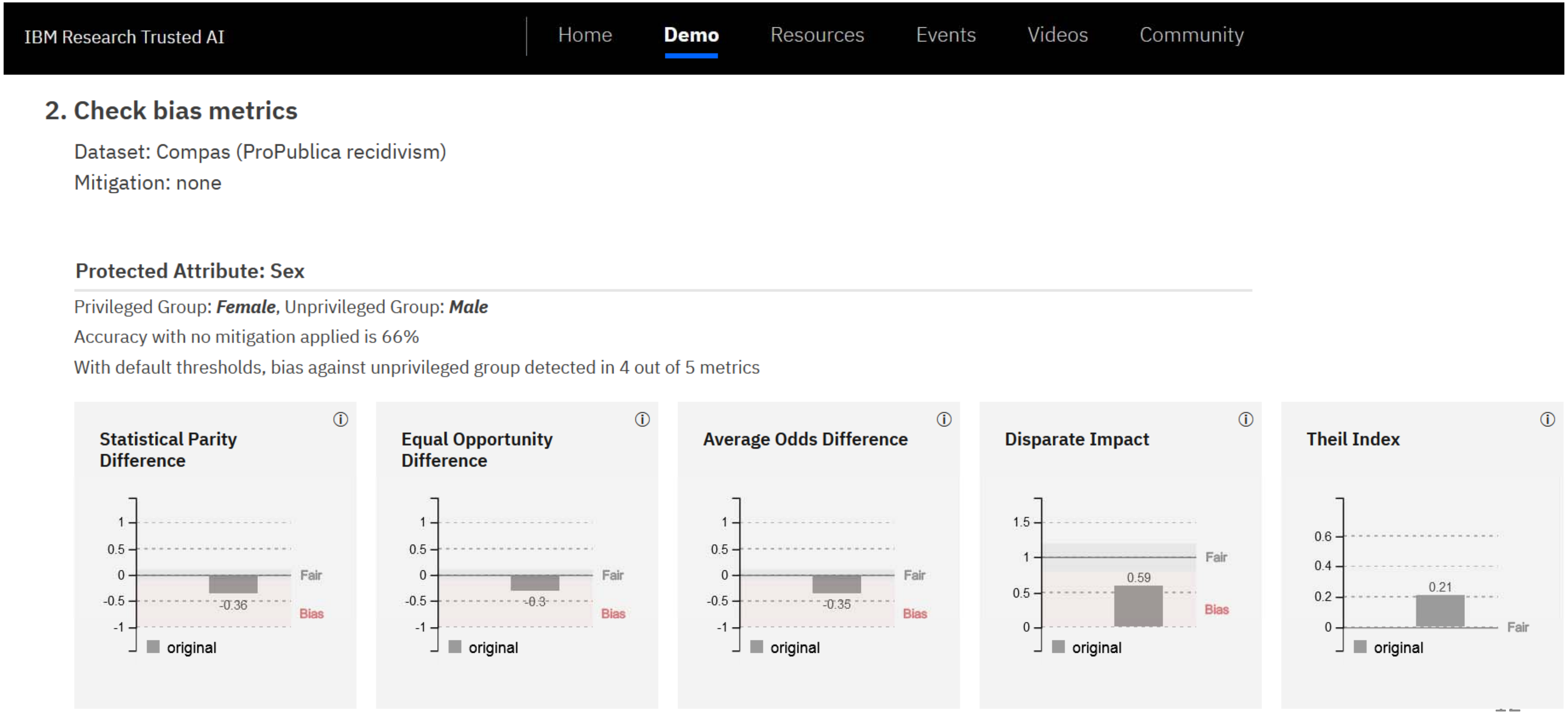
- https://www.data.gov/
  - US-centric agriculture, climate, education, energy, finance, health, manufacturing data, etc.
- https://datos.gob.es/es/catalogo
  - Spanish datasets in different domains
- https://cloud.google.com/bigquery/public-data/
  - BigQuery (Google Cloud) public datasets (bikeshare, GitHub, Hacker News, Form 990 non-profits, NOAA, etc.)
- https://www.kaggle.com/datasets
  - Microsoft-owned, various (Billboard Top 100 lyrics, credit card fraud, crime in Chicago, global terrorism, world happiness, etc.)
- https://aws.amazon.com/public-datasets/
  - AWS-hosted, various (NASA, a bunch of genome stuff, Google Books n-grams, Multimedia Commons, etc.)

2. Identify conflicts among selected datasets: Example

- **Dataset 1**: Data Related Jobs in US
  - https://www.kaggle.com/datasets/mohamedsiika/data-related-jobs-in-us
    - job title
    - seniority
- **Dataset 2**: IBM Employee Dataset
  - https://www.kaggle.com/datasets/rohitsahoo/employee
    - Job Role
    - Job Level

3. Be aware about bias and fairness in your datasets: Example (Identifying Bias)

- **AI Fairness 360**
  - https://aif360.mybluemix.net/
  - https://aif360.mybluemix.net/resources#tutorials



2. Check bias metrics

Dataset: Compas (ProPublica recidivism)
Mitigation: none

Protected Attribute: Sex

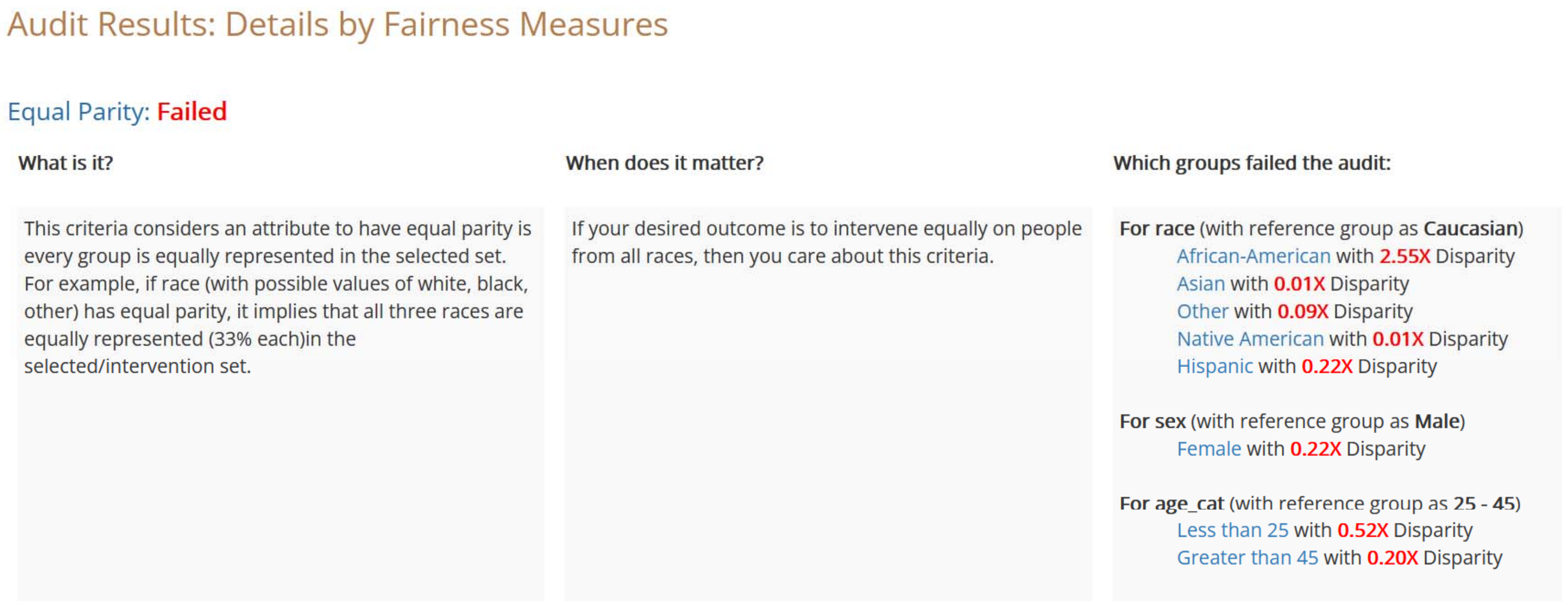
Privileged Group: ***Female***, Unprivileged Group: ***Male***

Accuracy with no mitigation applied is 66%

With default thresholds, bias against unprivileged group detected in 4 out of 5 metrics

3. Be aware about bias and fairness in your datasets: Example (Identifying Fairness)

- **Aequitas**
  - http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/



Audit Results: Details by Fairness Measures

Equal Parity: Failed

What is it?

This criteria considers an attribute to have equal parity is every group is equally represented in the selected set. For example, if race (with possible values of white, black, other) has equal parity, it implies that all three races are equally represented (33% each)in the selected/intervention set.

When does it matter?

If your desired outcome is to intervene equally on people from all races, then you care about this criteria.

Which groups failed the audit:

For race (with reference group as **Caucasian**)
African-American with **2.55X** Disparity
Asian with **0.01X** Disparity
Other with **0.09X** Disparity
Native American with **0.01X** Disparity
Hispanic with **0.22X** Disparity

For sex (with reference group as **Male**)
Female with **0.22X** Disparity

For age_cat (with reference group as **25 - 45**)
Less than 25 with **0.52X** Disparity
Greater than 45 with **0.20X** Disparity

General Instructions

- This assignment should be performed **in groups** composed of 2/3 students.
  - Students can also decide to perform this assignment in an individual way
- **Deadline**: Wednesday 12th April 2023
- The **assignment delivery** should include a PDF file
  - describing the main outcomes for each assignment step: selected datasets, conflicts identified, and reports from tools for bias and fairness
  - the main decisions taken (with respect to tools for bias and fairness)
  - the lessons learned and the difficulties found
  - and any other comments and suggestions

Course: Information Extraction, Retrieval and Integration

Unit 4: Data Integration

Assignment Description: Data Integration, Bias and Fairness

Mari Carmen Suárez de Figueroa Baonza
March 2023
Universidad Politécnica de Madrid