Gender Discrimination in Data Analysis: a Socio-Technical Approach

Supervisor: Prof. Viola Schiaffonati
Co-supervisor: Prof. Letizia Tanca
Prof. Pierre Senellart
Prof. Karine Gentelet

M.Sc. Thesis by:
Riccardo Corona
927975
07/10/2021

Data analysis

Set of processes for inspecting, cleaning, transforming, and modeling data with the aim of discovering useful information, informing conclusions, and supporting decision making.

Gender discrimination

Specific (sub)category of social problems, here expressed in the form of the so-called '**gender gap**', definable as:

> *A difference between the way men and women are treated in society, or between what men and women do and achieve.*

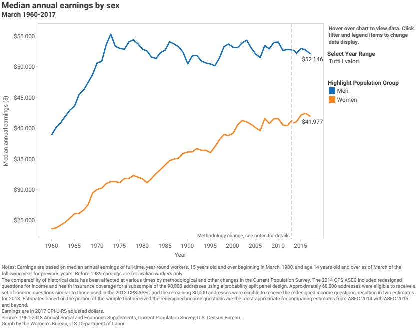
Scenarios & Problem Statement

Problem

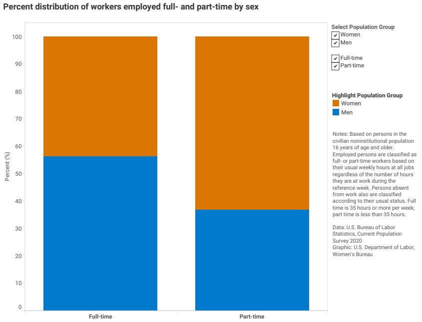
Data and datasets, on which a lot of actions of our daily routine are based, can be **unfair**. Unfair, or better to say, **biased** data, may influence, directly or indirectly, our perception of reality, and lead us to make decisions that, although seemingly fair and just, contain in turn bias, and discriminate against individuals or groups of individuals.

Example scenarios

- COMPAS tool used in the U.S. to predict recidivism risk biased against Black people (2016).
- Amazon software to screen candidates for employment biased against women (2015).

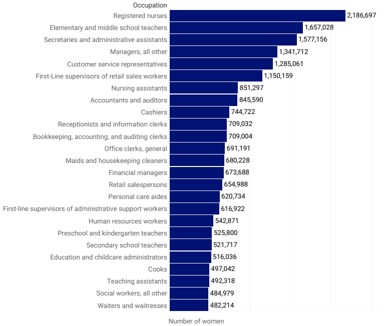


Median annual earnings by sex
March 1960-2017
Hover over chart to view data. Click filter and legend items to change data display.
Select Year Range
Tutti i valori
Highlight Population Group
Men
Women
$55.000
$50.000
$45.000
$40.000
$35.000
$30.000
$25.000
Median annual earnings ($)
$52.146
$41.977
Methodology change, see notes for details
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
Year
Notes: Earnings are based on median annual earnings of full-time, year-round workers, 15 years old and over beginning in March, 1980, and age 14 years old and over as of March of the following year for previous years. Before 1989 earnings are for civilian workers only.
The comparability of historical data has been affected at various times by methodological and other changes in the Current Population Survey. The 2014 CPS ASEC included redesigned questions for income and health insurance coverage for a subsample of the 98,000 addresses using a probability split panel design. Approximately 68,000 addresses were eligible to receive a set of income questions similar to those used in the 2013 CPS ASEC and the remaining 30,000 addresses were eligible to receive the redesigned income questions, resulting in two estimates for 2013. Estimates based on the portion of the sample that received the redesigned income questions are the most appropriate for comparing estimates from ASEC 2014 with ASEC 2015 and beyond.
Earnings are in 2017 CPI-U-RS adjusted dollars.
Source: 1961-2018 Annual Social and Economic Supplements, Current Population Survey, U.S. Census Bureau.
Graph by the Women's Bureau, U.S. Department of Labor



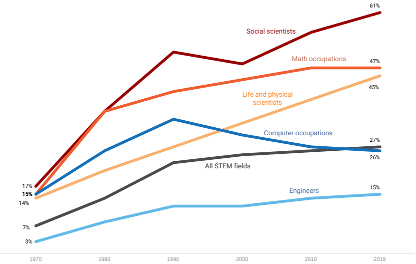
Percent distribution of workers employed full- and part-time by sex
100
90
80
70
60
50
40
30
20
10
0
Percent (%)
Full-time
Part-time
Select Population Group
Women
Men
Full-time
Part-time
Highlight Population Group
Women
Men
Notes: Based on persons in the civilian noninstitutional population 16 years of age and older. Employed persons are classified as full- or part-time workers based on their usual weekly hours at all jobs regardless of the number of hours they are at work during the reference week. Persons absent from work also are classified according to their usual status. Full time is 35 hours or more per week; part time is less than 35 hours.
Data: U.S. Bureau of Labor Statistics, Current Population Survey 2020
Graphic: U.S. Department of Labor, Women's Bureau

Most Common Occupations for Women in the Labor Force



Note: Full-time, year-round civilian employed 16 years and older. Occupations with at least 100 sample observations.
Data: U.S. Census Bureau, American Community Survey 2019
Graphic: U.S. Department of Labor, Women's Bureau

Percentage of science, technology, engineering, and math (STEM) workers who are women



Note: STEM occupations are classified according to the Standard Occupational Classification STEM recommendations for presentation of government data available at: https://www.bls.gov/soc/Attachment_C_STEM_2018.pdf
Source: U.S. Census Bureau, decennial census 1970-2000 and American Community Survey public use microdata 2010 and 2019.
Graphic by the Women's Bureau, U.S. Department of Labor

Tools for Assessing Fairness

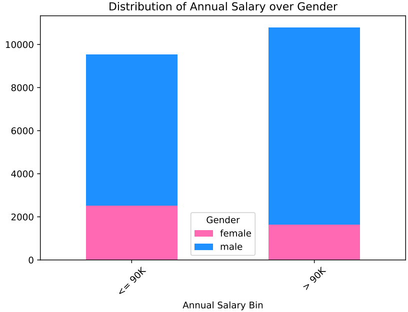
- *The 'Glassdoor Method'*: a framework for evaluating gender pay gap which relies on **linear regression**.
- *FAIR-DB*: an algorithm to detect bias in data based on **functional dependencies** and the related evaluation metrics.
- *Ranking Facts*: an application built on the idea of **ranking** which makes use of three statistical measures to evaluate fairness.

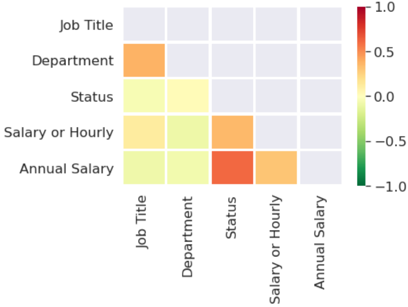
Case Study 1: Chicago

- **Data Preprocessing**: 20,309 tuples, of which 16,146 males and 4,163 females, and with 35 distinct *Job Title* values and 20 distinct *Department* values.

- **The 'Glassdoor Method'**: 24.2% 'unadjusted' pay gap; 0.4% 'adjusted' pay gap → no evidence of a systematic gender pay gap.
- **FAIR-DB**: 6 final functional dependencies; 11.4% of the dataset 'problematic' → dataset quite fair.
- **Ranking Facts**: dataset fair for both males and females, for each statistical measure.

Case Study 1: Chicago

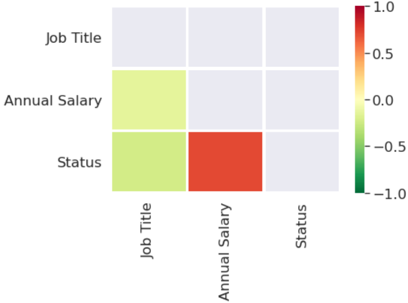




- **Data Preprocessing**: 22,996 tuples, of which 13,688 males and 9,308 females, and with 81 distinct *Job Title* values.

- **The 'Glassdoor Method'**: 30.4% 'unadjusted' pay gap; -0.5% 'adjusted' pay gap → no evidence of a systematic gender pay gap.
- **FAIR-DB**: 10 final functional dependencies; 24.3% of the dataset 'problematic' → dataset quite fair because of the low values of *difference* ('unfairness level') and *support* (number of tuples involved), but for higher-paying jobs men seem to have an economic advantage over women.
- **Ranking Facts**: dataset fair for males and unfair for females, for each statistical measure → proportion of women in the top-k ranking effectively very low.

Case Study 2: San Francisco



Other Design Choices

- **Part-time employees removal**: most of the tuples removed related to women (Chicago); excessive amount of tuples removed (San Francisco).
- **FAIR-DB: discretization using more bins**: less and different final dependencies detected (Chicago and San Francisco).
- **FAIR-DB: choice of different dependencies**: 85.6% (Chicago) and 92.5% (San Francisco) of the dataset 'problematic'.
- **Grouping of job titles**: overturning of the outcomes for Ranking Facts (Chicago dataset unfair for males and fair for females, for each statistical measure).
- **Voluntary introduction of bias**: results from each tool oriented toward unfair Chicago dataset, in which women are discriminated against (retaining 50%, 75%, and 90% of the *Annual Salary* value of female employees).

Section 1

Slide 1

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Subsection 1.1

This frame has an empty title.

- item 1
  - item 1.1
  - item 1.2
- item 2
- item 3

Slide 1.2 without numbering

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Block

Text.

Slide 2

Block

Text.

Alert block

Alert text.

Slide 2

Block

Text.

Alert block

Alert text.

Example block

Example text.