POLITECNICO DI MILANO
Master of Science in Computer Science and Engineering
Dipartimento di Elettronica, Informazione e Bioingegneria

Gender Discrimination in Data Analysis

Internal supervisor: Prof. Viola Schiaffonati
Internal co-supervisor: Prof. Letizia Tanca
External supervisor: Prof. Pierre Senellart
External co-supervisor: Prof. Karine Gentelet

M.Sc. Thesis by:
Riccardo Corona, 927975

Academic Year 2020-2021

To my friends – the longtime ones, those I have met at Politecnico, or around the world – for making my university life enjoyable and making me emotional in the moments we spent together.

To my parents, for supporting me always and by any means in my choices.

To my grandparents, for making me who I am today.

Ai miei amici – di vecchia data, conosciuti al Politecnico, o in giro per il mondo – per aver reso piacevole la mia vita universitaria e avermi fatto emozionare nei momenti trascorsi assieme.

Ai miei genitori, per avermi supportato sempre e con ogni mezzo nelle mie scelte.

Ai miei nonni, per avermi reso ciò che sono oggi.

Abstract

The abstract is a small summary of the thesis. It tells the reader in few words (up to one/one and a half page of total text) everything he/she needs to understand:

- ☐ the *context* of the work (e.g., chatbots),
- ☐ the specific *problem* approached by the thesis (e.g., the development of personal bots by non-programmers),
- ☐ if applicable, clearly state the *research questions* you would like to answer (e.g., "is it possible to enable non-programmers to do X using A?"),
- ☐ the three/four *core aspects of the proposed solution* (e.g., use pre-defined rules, use machine learning, assisted development, etc.),
- ☐ the *concrete outputs* produced by the thesis (e.g., a state of the art analysis, a conceptual/mathematical model, an application, middleware or API, an empirical study with/without users, etc.), and
- ☐ the *findings and conclusions* that one can draw from the evaluation of the approach (e.g., that under some very specific conditions non-programmers are indeed able to implement own chatbots effectively using the proposed technique).

Checklists
Now and there I propose checklists with items, such as the one just above this box. They are meant for you to check if you included all the content that is relevant and that should be included, in order to make your text complete. When reading your thesis, I will look for all these items.

Writing style

This is a M.Sc. thesis. It's neither Facebook nor Twitter nor an email. This is going to be an official document with legal value that will decide on the final mark of your yearlong university career and perhaps even on your future work perspectives. So, you surely don't want to be judged badly because of grammar errors, flawed/wrong vocabulary or superficial layout and/or text structure. It is a must that what you write is always *correct* content- and language-wise (no false statements or claims, no language mistakes), *readable* (no sentences that cannot be understood) and targeted at the *average-skilled reader* (professors, but also your own colleagues).

Plagiarism

This is a M.Sc. thesis. It's neither Facebook nor Twitter nor an email. This is going to be an official document with legal value that will decide on the final mark of your yearlong university career and perhaps even on your future work perspectives – yes, I plagiarized myself here a little bit. So, you surely don't want to copy/paste material from scientific articles, online resources, books, and similar without adequately acknowledging the holders of the respective intellectual property rights. If you do so, it is a must that you properly *cite* each source where you take text or inspiration from. It is fine to do so – actually, citing someone is a compliment! – but it becomes a crime if the source is not cited. Not only M.Sc. titles but also Ph.D. titles have been withdrawn for fraudulent "reuse" of others' intellectual property. Be aware that Politecnico di Milano, like most higher educational institutions that issue university degrees or scientific publishers, may use specialized software to automatically detect plagiarism.

Sommario

Here goes the translation into Italian of the abstract. If the thesis is written in Italian, no translation into English is needed. Hence, one of the following must be checked:

☐ Thesis written in *English*, properly proofread translation needed

☐ Thesis written in *Italian*, no translation needed, chapter omitted

Acknowledgements

Throughout the writing of this dissertation (and my whole university career) I have received a great deal of support and assistance.

I would first like to thank my external supervisor, Professor Pierre Senellart, whose expertise and constant support have been fundamental in sharpening my thinking and bringing my work to a higher level. In addition to his great professionalism and precision, the humanity that characterizes him has been of great help in moments of uncertainty, in which he has always been able to address me at best, making me look at the glass as half-full. It is a quality not to be taken for granted at all.

I would like to acknowledge my internal supervisor and co-supervisor, Professors Viola Schiaffonati and Letizia Tanca, always very kind and helpful, who have been able to effectively support me remotely in a notoriously not easy time, providing me with essential advice and feedback.

I would also like to thank my external co-supervisor, Professor Karine Gentelet, who unfortunately I never had the opportunity to meet in person for reasons of force majeure, but whose precious contribution has enriched this research work, allowing it to get out of the ordinary and to include interesting and diversified perspectives.

In addition, I would like to acknowledge my friends for having brightened my days, being close to me in difficult times, and having shared with me many meaningful experiences of my life and several unforgettable moments.

Finally, I would like to thank my parents and my grandparents, Angelo and Palma, who have always supported me throughout these difficult years, not only economically but also emotionally, pushing me not to give up and pursue new life experiences that may have enriched me, even knowing that those experiences, in some cases, would have taken me far from them.

Contents

Chapter 1

Introduction

1.1 Research Context: Gender Discrimination in Data Analysis

This research would like to be a bridge between two disciplines of very different nature, and which at first sight would seem to have little to do with each other: *data science* and *sociology*. As for the former, our focus will be on *data analysis*, that is, the set of processes for inspecting, cleaning, transforming, and modeling data with the aim of discovering useful information, informing conclusions, and supporting decision making. For what concerns the latter, we will focus on *social justice*, and even though most of the definitions will be provided in further chapters, it is appropriate to clarify the very fundamental concepts behind our research, starting from the notion of 'social problem'.

> *Social problem is a generic term applied to a range of conditions and aberrant behaviors which are manifestations of social disorganization. It is a condition which most people in a society consider undesirable and want to correct by changing through some means of social engineering or social planning.* [37]

In our research, we will focus on a specific category of social problems, namely those related to discrimination, and in particular on *gender discrimination*, expressed in the form of the so-called '**gender gap**'. According to [18], gender gap is definable as:

> *A difference between the way men and women are treated in society, or between what men and women do and achieve.* [18]

Specifically, the focus of our experiments will be *gender pay gap*, that is, the average difference between the remuneration for men and women in the

workforce, or, in other words, a measure of what women are paid relative to men. Our experiments will be centered on the economical perspective because it is the easiest to be measured in the data, being quantifiable for example as a number representative of a person's average monthly salary, but it is important to keep in mind that other facets of the gender gap problem come into play when dealing with sociological studies.

1.2 Scenarios & Problem Statement

Nowadays, digital devices have become pervasive in every aspect of our daily lives and almost every action we perform leaves a digital trail. We generate data whenever we go online, when we communicate with people through any kind of application, when we shop, or even just when we carry our smartphones. It is therefore of societal and ethical importance to ask whether data and datasets, on which so many actions of our daily routine are based, are *fair* or not. Unfair, or better to say, *biased* data, may in fact 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 therefore discriminate against individuals or groups of individuals. Without going into too much detail, which will be deepened in the following chapters, we will now provide the reader with a couple of scenarios that can intuitively give the idea of the problem.

A first example, related to racial discrimination, is given by [35]. A study conducted by the University of California in 2019 concluded that an algorithm used to allocate health care to patients in U.S. hospitals was less likely to refer black people than white people who were equally sick to programs that aim to improve care for patients with complex medical needs. Although there was no discriminatory intent, misconceptions in the design phase led to the introduction of bias, and consequently to the involuntary discrimination of millions of black citizens.

Another example, more closely linked to the focus of our research, namely gender discrimination, is provided in [30]. The article reports the outcomes of a study led by the University of Southern California researchers in 2021, who found that Facebook systems were more likely to present job ads to users if their gender identity mirrored the concentration of that gender in a particular position or industry. Specifically, the team of researchers bought ads on Facebook for delivery driver job listings that had similar qualification requirements but for different companies: Domino's, who has more male drivers, and Instacart, who has more female ones. The study found that Facebook targeted the Instacart delivery job to more women and

the Domino's delivery job to more men. Similar findings were obtained by testing software engineer job listings for Nvidia and Netflix, and therefore the researchers, in their conclusions, spoke of "a platform whose algorithm learns and perpetuates the existing difference in employee demographics".

Further examples and explanations will be provided in the following chapters, but what we want to clarify is that our goal is not to solve the problem of discrimination in data, which is a huge and multifaceted issue which encompasses a human dimension very difficult to address, but rather to take a look at the current state of the art, observing some tools in action and trying to highlight their strengths and weaknesses, and also providing a non-technical perspective to give a broader picture of the situation. We believe that this research may represent a good starting point for those who work in the sector to have an overview of the problem and to better understand what to focus on.

1.3 Methodology

We will approach the problem by first providing the reader with some preliminary socio-ethical and technical concepts taken from a *systematic literature review*, and later on we will describe, referring to the documentation at our disposal, the tools we decided to adopt for our analysis.

After that, we will conduct *parallel research* in sociology and information technology, distinguishing some *case studies* for our experiments and trying to interpret our results also looking at the sociological background introduced beforehand. The experiments themselves are basically *comparative studies* of the algorithms, and, in terms of *software instruments*, we will rely mainly on Python and Jupyter Notebook, with the various packages and libraries made available by developers.

1.4 Thesis Structure

We provide here a brief overview of the contents of each chapter, in order to help the reader move through the thesis and find specific information. The beginning of each chapter provides more detailed descriptions about the contents of the chapter itself.

- Chapter 2: **Socio-Ethical Preliminaries**. The aim of this chapter is to provide the reader with preliminary notions of ethical and sociological (rather than technical) nature.

- Chapter 3: **Technical Preliminaries**. The aim of this chapter is to provide the reader with preliminary notions about the technical knowledge necessary to understand the functioning of the adopted tools, which will be described later on in Chapter 5.
- Chapter 4: **Sociological Research**. The aim of this chapter is to provide the reader with a sociological background by reporting information about gender gap in the society.
- Chapter 5: **Techniques**. The aim of this chapter is to provide the reader with an overview on the tools adopted in our experiments.
- Chapter 6: **Experiments**. The aim of this chapter is to describe the experiments conducted using the tools introduced in Chapter 5, in order to verify the presence (and the nature) of bias in our datasets and discuss their fairness, according to the concepts introduced in Chapter 2 and the sociological background explored in Chapter 4.
- Chapter 7: **Conclusions & Future Work**. The aim of this chapter is to draw the conclusions of our research, mixing the results obtained in our experiments described in Chapter 6 with the sociological background depicted in Chapter 4, and recalling some preliminaries exposed in Chapter 2 and Chapter 3 when needed.

Chapter 2

Socio-Ethical Preliminaries

The aim of this chapter is to provide the reader with preliminary notions of ethical and sociological (rather than technical) nature.

Starting from the concept of *bias*, passing through *discrimination* and *human rights*, we will discuss about *equality*, *equity* and finally *fairness*, by providing definitions and relevant examples from the literature on these topics, with a focus on the data and computer systems perspective. It is important to emphasize that, despite the exhaustiveness of the definitions, these terms often have different meanings depending of the context of use, and there is a lot of debate on how to interpret them and eventually include all their dimensions in computer systems.

Because of the 'dual nature' of this research, this chapter has to be seen as complementary to Chapter 3, in which some technical bases necessary to understand the functioning of the adopted tools will be provided.

2.1 Bias

Although the word '**bias**' does not have an intrinsically negative meaning (it is informally used to indicate a deviation from neutrality), it is mostly adopted in contexts where it entails a moral and social dimension. As reported in [25]:

> *We use the term bias to refer to computer systems that* systematically *and* unfairly discriminate *against certain individuals or groups of individuals in favor of others. A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate.* [25, p. 332]

Therefore, it is important to underline that unfair discrimination due to bias is strictly related to systematic and unfair outcome, where the word 'systematic' is used with the meaning of 'regular, which occurs methodically when certain conditions arise'.

By following the classification provided in [25], we can distinguish three overarching categories of bias:

- **Preexisting bias**: it has its roots in social institutions, practices and attitudes. Preexisting bias may originate in the society at large or in subcultures and organizations (*societal bias*), but it is also intrinsic in the nature of every human being (*individual bias*), and can enter a computer system either voluntarily or implicitly and unconsciously, even in spite of the best intentions of the system designer. Furthermore, since preexisting bias is often related to historical discrimination of disadvantaged groups, it could lead to the introduction or the exacerbation of representation issues in the data. An example of preexisting (gender) bias is the one present in the society that leads to the development of educational software that overall appeals more to boys than girls [25].

- **Technical bias**: it arises from the resolution of issues in the technical design. Technical bias may originate from design choices, constraints, and technological tools, or exacerbate preexisting bias. An example of technical bias, due to technical constraints, is the one of a monitor screen displaying the flight options most relevant to an airline customer: the screen dimension forces a piecemeal representation of the flights and therefore if the ranking algorithm systematically places certain flights on initial screens and other flights on later screens, it exhibits technical bias [25].

- **Emerging bias**: it emerges some time after a design is completed, as a result of changing societal knowledge, population, or cultural values. Emerging bias is strictly related to the specific context of use, and it is the most difficult to detect. An example of emerging bias, caused by a *mismatch between users and system design* due to *different values* (that is, originated when a computer system is used by a population with different values than those assumed in the design), is the one of an educational software embedded in a game situation that rewards individualistic and competitive strategies used by students with a cultural background that eschews competition and promotes collaboration [25].

For the purpose of this research, we will focus on preexisting bias (in

particular, societal bias) and technical bias, but it is important to point out that emerging bias should not be underestimated in the long run, especially when it arises from a mismatch between users and system design due to different values, because society is in constant change and systems should be re-adjusted or re-invented in order to keep up with the present.

A significant example of preexisting (racial) bias, exacerbated by technical bias, is provided in [3]: a commercial tool called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) was used in courts in the U.S. to automatically predict some categories of future crime to assist in bail and sentencing decisions. On average, the tool correctly predicted recidivism 61% of the time, but blacks were almost twice as likely as whites to be labeled a higher risk but not actually re-offend. The tool made the opposite mistake among whites: they were much more likely than blacks to be labeled lower risk but go on to commit other crimes.

Other examples, related to gender bias and technology, are given by [27] and [15]. The former is about a research conducted in 2015, in which a tool was used to simulate job seekers that did not differ in browsing behavior, preferences or demographic characteristics, except in gender. One experiment showed that Google displayed adverts for a career coaching service for '$200K+' executive jobs 1,852 times to the male group and only 318 times to the female group. The latter concerns another Big Tech company, Amazon, whose machine learning specialists, back in 2015, discovered that their new recruiting engine was not rating candidates in a gender-neutral way, because the system taught itself that male candidates were preferable by penalizing resumes that included the word 'women's'.

2.2 Discrimination

Bias can lead to **discrimination**, but what discrimination is and how it occurs is a controversial issue. As reported in [43], often the law, rather than providing a definition of discrimination, defines a list of attributes, called **protected attributes**, that cannot be used to take decisions in various settings. The list is non-exhaustive and includes characteristics such as race, gender, religion, or sexual orientation. Groups of people that are more likely to be discriminated against because of these attributes are therefore classified as *protected groups*.

Trying to elaborate a bit more, we can define discrimination as the result of either one or both the following:

- **Disparate treatment**, or *intentional discrimination*: the illegal prac-

tice of treating an entity, such as a job applicant, differently based on a protected attribute such as race, gender, age, religion, sexual orientation or national origin because of a discriminatory motive.

- **Disparate impact**, or *unintentional discrimination*: the result of structural disparate treatment, in which policies, practices, rules or other systems that appear to be neutral result in a disproportionate adverse impact on a protected group. Disparate impact is not based on a discriminatory motive and the discriminating agent is usually unaware of the discrimination.

Protected attributes are mentioned in the article 2 of the Universal Declaration of Human Rights (UDHR), which states:

> *Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.* [4]

Discrimination is therefore strictly related to *human rights*, together with the concept of *equality*, since 'equal', 'equally' and 'equality' itself are recurring words in several articles of the Declaration.

2.3 Human Rights

For what concerns **human rights**, there is a lot of debate on how to incorporate them in computer systems by following a 'human-rights-by-design' approach [40], in order to contrast the negative effects of the so-called 'dual-use technologies': products which may serve legitimate societal objectives but are also used to undermine human rights like freedom of expression or privacy. Of course, reaching this goal would require a culture shift and huge efforts from both national governments and businesses, which should design tools, technologies, and services to respect human rights by default, rather than permit abuse or exploitation of them. A similar concept is proposed in [51], where the authors sketch the contours of a comprehensive governance framework for ensuring AI systems to be ethical in their design, development and deployment, and not violate human rights. This framework should be effective in contrasting *ethics washing*: the practice of fabricating or exaggerating a company's interest in equitable AI systems that work for everyone, a sort of side door that companies use to substitute regulation with ethics.

For the purpose of this research, we can define human rights as "inalienable fundamental rights to which a person is inherently entitled simply because she or he is a human being" [45, p. 3]. A few examples are the rights to life and liberty, freedom from slavery and torture, freedom of opinion and expression, the rights to work and education, and the right to the pursuit of happiness. These norms are concerning every human being, regardless of sex, age, language, religion, ethnicity, or any other status.

2.4 Equality & Equity

Equality is generally intended as "an ideal of uniformity in treatment or status by those in a position to affect either" [8]. The concept of equality is often associated with discrimination mostly because of the article 7 of the UDHR, which states:

> *All are equal before the law and are entitled without any discrimination to equal protection of the law.* [4]

This principle is known as 'equality before the law', and establishes that everyone must be treated equally under the law regardless of race, gender, color, ethnicity, religion, disability, or other characteristics, without privilege, discrimination, or bias.

However, it is important to distinguish between two different political and social theories:

- **Equality of opportunity**:

  > *The idea that people ought to be able to compete on equal terms, or on a "level playing field", for advantaged offices and positions.* [38]

  This principle is based on the notion of *sameness*, where fairness is achieved through equal treatment regardless of people's needs. Equality of opportunity is usually simply referred as **equality**, and from now on we will adopt the same terminology for this research.

- **Equality of outcome**: the idea that people should have access to resources (possibly of a different nature and to a different extent) in order to be able to reach the same condition. This principle is based on the notion of *need*, where fairness is achieved by treating people differently depending on their endowments and necessities. Equality of outcome is also known as **equity**, and from now on we will adopt the same terminology for this research.

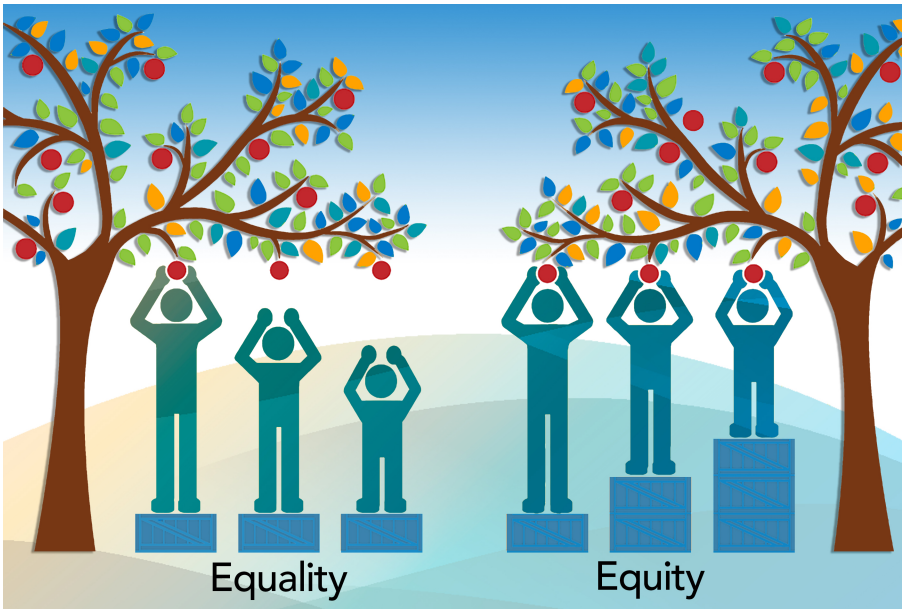


Figure 2.1: Visual example of the difference between equality and equity. ©2014, Saskatoon Health Region. Source: `https://www.nwhu.on.ca/ourservices/Pages/Equity-vs-Equality.aspx`.

Figure 2.1 provides a simple example of the difference between equality and equity. Treating people equally, in this scenario, means to give everyone the same one box to reach the fruit, while treating people equitably means to give them as many boxes as they need to achieve the goal. It is important to notice that equity could require (and often requires) unequal treatment.

Moving on to the data perspective, we can now extend (or restrict, depending on the point of view) the equality and equity concepts to data equality and data equity. Data equality usually refers to transparency of institutions and companies towards customers regarding the information collected on their account, whereas data equity is used in a different context. The authors of [33] distinguish between four different facets of data equity:

- **Representation equity**: bias may arise because of material deviations between the data and the world represented by the data, often with respect to historically disadvantaged and underrepresented groups. Even when dealing with contemporary data, disparities rooted in historical discrimination can lead to representation inequities and therefore to the introduction or the exacerbation of problems. For example, in the U.S. there has been a lot of discussion about racial disparities concerning COVID-19, regarding both availability of testing

(fewer test sites in minorities neighborhoods, historically poorer) and desire of individuals to be tested (black people more suspicious about the medical system, because of their history of unfair treatments) [33]. Another example, already mentioned in Section 2.1, is related to Amazon and a software developed by the company for screening candidates for employment: the software was trained on the already hired employees and since they were mostly males, females became underrepresented in the data and the software was much more likely to mark women as unsuitable for hiring.

- **Feature equity**: bias may arise because not all the features needed to represent a marginalized group of people and required for a particular analysis are available in the data, or because some of these features are voluntarily removed in the decision-making process. As an example, for a specific study involving transgender people, it may be important to distinguish between their birth name and their self-assigned name.

- **Access equity**: bias may arise because of a non-equitable and participatory access to data and data products across domains and levels of expertise due for instance to the opacity of data systems or the need to respect the privacy of data subjects. A classical example is the one of medical records: making them public could lead to the development of new techniques to eradicate diseases, but on the other side most people are very sensitive about sharing medical information because of the simplicity of re-identify anonymized data, and there are a lot of regulatory constraints on such sharing.

- **Outcome equity**: bias may arise because of a lack of monitoring and mitigation of unintended consequences for any group affected by the system after deployment, directly or indirectly (for example, contact tracing apps may facilitate stigma or harassment [32]).

2.5 Fairness

As discussed in Section 2.4, both equality and equity aim to achieve **fairness**, despite the different approaches of the two theories, but what fairness really is is a widely debated topic. A very generic definition, taken from [17], depicts it as "the quality of treating people equally or in a way that is right or reasonable".

In the sociological context, fairness is often seen as a synonym of *justice*, and consequently **social justice** is fairness as it manifests in the society,

described by [6, p. 405] as "an ideal condition in which all members of a society have the same rights, protections, opportunities, obligations, and social benefits". Although the literature on this subject does not always agree on their number, we can delineate five interrelated principles of social justice, by following the classification provided in [1]:

- **Access to resources**: a just society should provide services and resources that are available to each different socioeconomic group, in order to give everyone an equal start in life.
- **Equity**: in unjust societies, there are always disenfranchised groups. These groups need to receive more support from the society than privileged ones, in order to move towards the same outcome.
- **Participation**: everyone in a just society, and not just small groups of individuals, should be able to participate in the decisional processes that affect their lives.
- **Diversity**: a just society should recognize the value of diversity and cultural differences, and develop ad-hoc policies with the aim of breaking down societal barriers.
- **Human rights**: a just society should ensure the protection of everyone's civil, political, economic, cultural, and social rights.

Moving back to the data and computer systems perspective, and recalling the aforementioned concepts of equality and equity, we can distinguish between two different concepts of fairness [21]:

- **Individual fairness**: any two individuals who are similar *with respect to a task* should receive similar outcomes. The similarity between individuals should be captured by an appropriate metric function, usually difficult to determine. For example, deciding whether or not to display a specific advertisement is a classification task, and the definition of individual fairness assumes the existence of a task-specific metric (e.g. the number of clicks made by the users) capable of determining, for any two individuals, how (dis)similar they are for the specific task. Individual fairness is strictly related to the idea of equality.
- **Group fairness** (also known as *statistical parity*): demographics of the individuals receiving any outcome - positive or negative - should be the same as demographics of the underlying population. For example, in the problem of predicting whether or not to hire applicants, assuming

to divide them into groups according to their gender, this means the acceptance rates of the applicants from the groups must be equal regardless of the protected attribute. Group fairness equalizes outcomes across protected and non-protected groups, and is therefore strictly related to the idea of equity.

Although individual and group fairness are not mutually exclusive in theory, in real life it is often hard to conciliate the two approaches. Furthermore, this categorization is not the only possible one: the authors of [48] collected and provided about twenty among the most prominent definitions of fairness, and applied each of them to a case study based on gender-related discrimination, in which the aim was to assign a credit score to people requesting a loan by using *Personal status and gender* as a protected attribute for the decision-making process, operated by a classifier (an algorithm that automatically orders or categorizes data into one or more of a set of 'classes', in this case only 'good credit score' and 'bad credit score').

Among the others, a couple of peculiar definitions, often listed together with individual and group fairness, are the following:

- **Fairness through unawareness**: protected attributes are not used in the decision-making process, and therefore the subsequent decisions cannot rely on them. This 'blind' approach relies on *impartiality* and is consistent with the disparate treatment principle, but removing features means losing information, and furthermore there could be features correlated to protected attributes that would not be removed, potentially introducing bias.

- **Counterfactual fairness**: a precise and non-technical definition is provided in [49]:

  > *A model is fair if for a particular individual or group its prediction in the real world is the same as that in the counterfactual world where the individual(s) had belonged to a different demographic group. However, an inherent limitation of counterfactual fairness is that it cannot be uniquely quantified from the observational data in certain situations, due to the unidentifiability of the counterfactual quantity.* [49, p. 1]

  To better clarify the concept, we could imagine a situation in which a software has the task of deciding whether or not to assign a promotion to the employees of a company by looking at their profile that includes,

among the other attributes, sex and race. The software is counterfactually fair if, for each individual, the outcome of the analysis is the same both in the case in which the real values of these attributes are used and in the case in which these values are replaced with others (counterfactuals).

The classifier resulted to be fair depending on the notion of fairness adopted, showing the impossibility of addressing fairness as a unique, broad and inseparable concept. This result is coherent with *Chouldechova's impossibility theorem* [13], which demonstrates, taking three definitions of fairness, the impossibility of satisfying all of them.

Chapter 3

Technical Preliminaries

The aim of this chapter is to provide the reader with preliminary notions about the technical knowledge necessary to understand the functioning of the adopted tools, which will be described later on in Chapter 5.

We will start by introducing *relational databases*, the *data science pipeline* and *data mining techniques*, and then focusing on more specific concepts such as *linear regression* and *functional dependencies*. Finally, an explanation of some needed *evaluation metrics* and *statistical concepts* will follow.

As specified in Chapter 2, the complementarity of the preliminaries is one of the key points of this research, and will allow the reader to have two perspectives of a different nature on the same problem.

3.1 Relational Databases

When dealing with computer systems, one of the most basic notions, often inappropriately taken for granted, is the one of '**data**', definable as:

> *Information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer.* [16]

Therefore, a large amount of data stored in a computer in some organized manner is called a **database**. To be more precise, a database is "any collection of data, or information, that is specially organized for rapid search and retrieval by a computer" [9]; while the software that supports the management of these data is called a **Database Management System (DBMS)**.

The history of databases is deeply interconnected with the history of computer science itself, because the problem of how to store and retrieve

information appeared as one of the initial challenges of computer creators. However, in the past few decades the rapid and enormous evolution of computer systems and databases led to the adoption and the development of the so-called 'data models'. A **data model** [2] is an abstract representation of an information system, which defines the data elements and the relationships between data elements. The aim of a data model is to give a clear and intuitive overview on how a system looks like, by providing a standardized description of its components, in such a way as to facilitate the understanding of the system itself and the possible integration with other systems.

Nowadays, the most widespread data model is the **relational model**, firstly proposed by Codd in [14]. The relational model represents a database as a collection of relations, depicted as tables of values. Each row of the table is a collection of related data values, referring to a real-world entity or relationship between entities. Therefore, we can simply define a **relational database** as a digital database based on the relational model of data. To make it clearer, the following list provides the main terms used in this context, together with a concise explanation, while Figure 3.1 shows them in a trivial example.

- **Table**, or **relation**: modeling of a real-world entity or of a relationship between real-world entities.
- **Row**, or **tuple**: single data record.
- **Column**, or **attribute**: property, or feature, of a relation.
- **Cardinality**: total number of tuples of a relation.
- **Degree**: total number of attributes of a relation.
- **Primary key**: attribute, or combination of attributes, that uniquely identifies a tuple among the others.
- **Domain**, or **data type**: set of values that a specific attribute can assume (for example, integer numbers, or boolean values).
- **Database schema**, or simply **schema**: blueprint of the database that outlines the way its structure organizes data into tables.
- **Database instance**, or simply **instance**: set of tuples in which each tuple has the same number of attributes as one of the relations of the database schema. It specifies the actual content of the database.
- **Integrity constraint**: property that is supposed to be satisfied by all instances of a database schema.

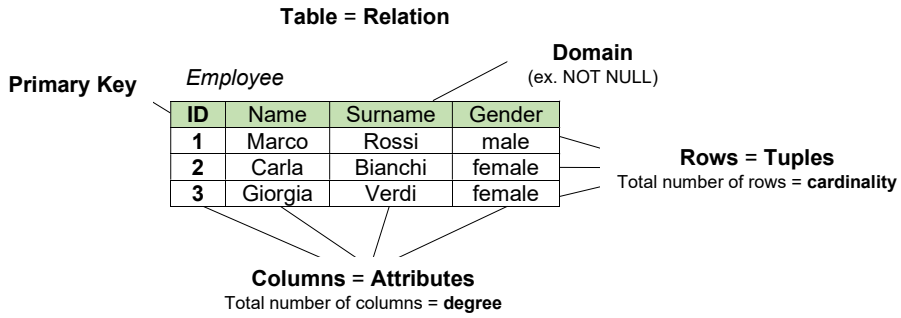


Figure 3.1: Relational model concepts in a trivial example. 'Employee' is the name of the real-world entity of reference and therefore of the related table in the model.

Lastly, since this term will be often used in the subsequent sections and chapters, we define a **dataset** as a collection of data. More specifically, since our data are in a tabular format according to the relational model, a dataset simply corresponds to one or more database tables.

Further details on relational databases can be found in [2].

3.2 Data Science Pipeline

Because of the broadness of the concept, there is not a unique and precise definition of data management. In general, we can identify it as the process of acquiring, storing, organizing, and maintaining data created and collected by an organization. In [26], the author, referring to [34], classifies *data management*, together with *analytics*, as one of the two sub-processes to extract insights from data, while the overarching process is referred as **data science pipeline**, or *big data pipeline*. For the sake of clarity, since the term is the one used in [26], we define big data as:

> *Large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.* [39, p. 10]

However we preferred to adopt the name of 'data science pipeline' instead of 'big data pipeline', since we will not deal with big data, which are not a concept strictly inherent to this research.

Since fairness should be addressed in each phase of the data science pipeline, the subsequent list provides a concise explanation of the operations performed in each step, by following the classification proposed in [31], together with the main potential sources of bias.

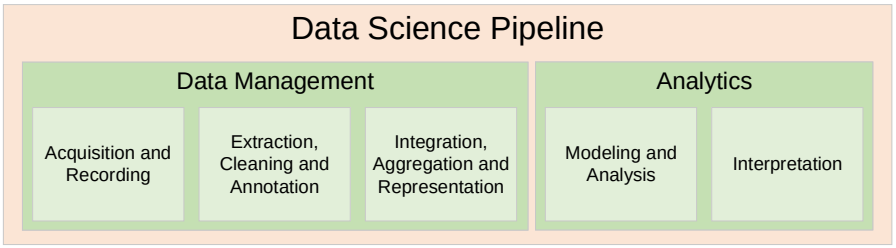


Figure 3.2: Data science pipeline. Image based on the one shown in [26].

- **Acquisition and recording**: data are recovered and captured. In this phase the introduction of bias could derive from some preliminary critical choices we have to deal with, concerning the availability of sources, the identification of who is represented by the data, the definition of what has been measured and of our duties to the people in the data (for example, we may owe them a certain degree of privacy).

- **Extraction, cleaning and annotation**: real data are most of the time messy and dirty, therefore we need to extract the relevant information and clean them, in order to express them in a structured form suitable for analysis. Unfortunately, data cleaning itself is based on assumptions, and wrong assumptions may lead to bias (for example, we may assume missing values in the data as missing at random, while there could be other, maybe ethical, reasons behind).

- **Integration, aggregation and representation**: data analysis often requires the collection of heterogeneous data from different sources, therefore we need to integrate them in order to guarantee syntactic and semantic coherence. Again, we have to rely on assumptions on the world, as for the case of data representation, in which a lot of choices are made in order to decide what to represent, potentially leading to bias (for example, in the context of sentiment analysis we may ascribe sentiment to labels, or we may decide to group age values instead of considering every single year).

- **Modeling and analysis**: before the actual analysis, an abstract model of the data is generated, in order to capture the essential components of the system and their interactions. However, the process of abstraction of concrete data in a conceptual standard model necessarily leads to the loss of information (for example, the relational model provides an intuitive overview of the system, but it does not include any semantics).

- **Interpretation**: a decision-maker, provided with the results of the analysis, has to interpret these results. This process usually requires to examine all the assumptions made and to retrace the analysis, and because of the complexity of the task and the problems that may arise from computer systems (bugs, errors), a human (and therefore impossibly perfectly fair) supervision is needed (for example, the failures of system components can go unnoticed and result in loss of data, or the data format may have changed without being notified, and therefore the system should be equipped with monitoring scripts and mechanisms to obtain user confirmation and correction).

3.3 Data Mining Techniques

> *Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases.* [46, p. 80]

Data mining is a broad topic, and usually a variety of procedures are needed in order to gain knowledge from data. However, we can distinguish three main categories of techniques, in each of which the fairness problem should be addressed differently:

- **Preprocessing techniques**: procedures used to transform the raw data in a useful and efficient format. The aim is to improve the overall quality of the data and consequently the data mining results.
- **Inprocessing techniques**: data are subjected to various methods using machine learning and artificial intelligence algorithms to generate a desirable output.
- **Postprocessing techniques**: methods to evaluate the extracted knowledge, visualize it, or merely document it for the end user. The knowledge can also be interpreted and incorporated into an existing system.

For the purpose of this research, we will focus on preprocessing techniques, which constitute one of the most critical steps in the data mining process, since they deal with the preparation and transformation of the initial dataset; while the bias analysis we will perform making use of the tools adopted is part of data (in)processing. Data preprocessing methods are divided into four categories [46]:

- **Data cleaning**: since real-world data are often incomplete, noisy, and inconsistent, some routines are needed in order to fill in missing values, smooth out the noise and correct the inconsistencies. For what concerns *missing values*, these procedures include the removal of the specific tuple, or the filling (manual or automatic) of the missing value by using, for example, a constant (e.g. 'unknown'), the mean (for numerical attributes) or simply what is perceived to be the most probable value. Noise instead can be seen as a random error or variance in a measured variable, and some smoothing techniques for *noisy data* are:
  - **Binning**: a data value is smoothed by looking at its 'neighborhood', that is, the values around it.
  - **Regression**: data are fitted to a function, in order to be smoothed according to the function itself. A specific type of regression, useful for our analysis, is *linear regression*, which will be further explored in Section 3.4.
  - **Clustering**: similar data are organized into groups of values, called 'clusters'. Values that fall outside the set of clusters may be considered as outliers.
- **Data integration**: as mentioned in Section 3.2, data often come from different (possibly heterogeneous) sources, and therefore they need to be combined in order to obtain a coherent model and remove inconsistencies (such as redundancies between attributes, where some can be derived from others).
- **Data transformation**: data are transformed or consolidated in appropriate forms suitable for the mining process. Some techniques used in this context are:
  - **Normalization**: data values are scaled so as to fall within a specified range, such as $(-1.0, 1.0$ or $0.0, 1.0)$.
  - **Aggregation**: new attributes are constructed from the given set of attributes to help the mining process by summarizing or aggregating information (for example, daily sales data may be aggregated so as to compute annual total amounts).
  - **Generalization**: raw (or low-level) data are replaced by higher-level ones, by following a specific hierarchy (for example, the attribute *city* can be generalized to *country*).

  - **Discretization**: raw values of numeric attributes are replaced by interval levels or conceptual levels (for example, age values between 15 and 18 could be labeled as 'adolescence').

- **Data reduction**: in order to make mining more effective and get better analytical results, several techniques can be applied to obtain a reduced representation of the dataset that is much smaller in volume, yet closely maintains the integrity of the original data. These methods include, among the others, *attribute subset selection*, in which attributes considered as not particularly relevant for the analysis are removed, and *numerosity reduction*, where data are replaced by smaller data representations, such as parametric models.

3.4 Linear Regression

In order to fully understand how one of the adopted tools works (the one we refer to as 'Glassdoor Method', described later in Section 5.1), it is appropriate to have a closer look at **linear regression** [28]. As mentioned in Section 3.3, linear regression is a preprocessing technique used to smooth out noise or to find patterns within a dataset, which attempts to model the relationship between two or more variables by fitting data to a linear equation (represented, in the two-variable case, by a straight line in a Cartesian plane). The results from linear regression help in predicting an unknown value depending on the relationship with the predicting variables. For example, the height and weight of an individual generally are related: usually taller people tend to weigh more. We could use regression analysis to help predict the weight of a person, given their height.

We can distinguish between **simple linear regression**, in which a single input variable is used to model a linear relationship with the target variable (as for the example of height and weight), and **multiple linear regression**, where more predicting variables are used.

For simple linear regression, the reference equation is:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Variable x is called *explanatory* or *independent variable*, while y is referred to as *dependent variable*; β_1 is the *slope* of the line, also known as regression coefficient, and β_0 is the *intercept* (the value of y when $x = 0$), while ϵ is the *error* in the estimation of the regression coefficient, also known as residuals, which account for the variability in y that cannot be explained by the linear relation between x and y.

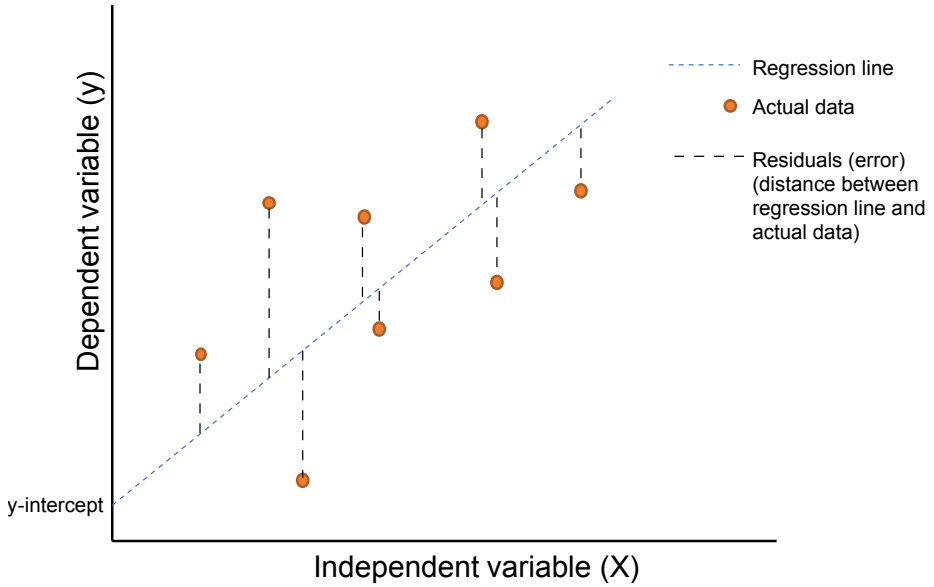


Figure 3.3: Simple linear regression graph.
Source: `https://www.reneshbedre.com/assets/posts/reg/mlr/residual.svg`.

For multiple linear regression, the formula is generalized in order to encapsulate also the other independent variables $(x_1, \ldots, x_n)$ and the related slope coefficients $(\beta_1, \ldots, \beta_n)$:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \epsilon$$

Figure 3.3 shows a simple linear regression graph. It is important to point out that the x and y variables remain the same, since they represent data features that cannot be changed, while the values that we can control are the slope and the intercept. Indeed, there can be multiple straight lines depending upon the values of intercept and slope, and what the linear regression algorithm does is to fit multiple lines on the data points and return the line that results in the least error.

Another important parameter for regression analysis is R^2, also known as *coefficient of determination* (or *coefficient of multiple determination* for multiple linear regression). It is a statistical measure of how close the data are to the fitted regression line, and therefore it indicates how much variation of the dependent variable is explained by the independent variable(s) in a regression model. R^2 values range from 0 to 1 and are commonly stated as percentages from 0% to 100%, where 0% refers to a model that explains none of the variability of the data around its mean, while 100% refers to a model that explains all the variability of the data around its mean (in this case, all

the actual data values would be on the regression line).

Formally, we can define R^2 as:

$$R^2 = 1 - \frac{UnexplainedVariation}{TotalVariation} = 1 - \frac{\sum_i (y_i - \hat{y}_i)}{\sum_i (y_i - \bar{y})}$$

where y_i is one of the actual data values, $\hat{y}_i$ is the corresponding predicted value, and $\bar{y}$ is the mean of all the y_i values, for $i = 1, \ldots, n$.

Generally speaking, at least for the purpose of this research, the higher the R^2 value, the better the model fits the data.

3.5 Functional Dependencies

One of the tools adopted, FAIR-DB, uses specific classes of integrity constraints, known as dependencies, to detect unfair behaviors in datasets (further details on the tool will be provided in Section 5.2). A **dependency** is a constraint that applies to or defines the relationship between attributes, and it occurs in a database when information stored in a table uniquely determines other information stored in the same table. Basically, dependencies are constraints not necessarily imposed by the system designer but intrinsically satisfied by the data.

Functional Dependencies (FDs) are a specific type of dependency, involving two (sets of) attributes of the same relation in which the first uniquely determines the second or, in other words, knowing the value of one attribute (or set of attributes) is enough to tell the value of the other one. The notation to indicate a functional dependency is:

$$A \rightarrow B$$

which can be read as 'B is functionally dependent upon A', or 'A uniquely determines B', whereas A and B are attributes (or eventually sets of attributes) of a table. A is called antecedent, or *left-hand-side (LHS)*, and B consequent, or *right-hand-side (RHS)*. An example of functional dependency could be the one of a table containing the information about the employees of a company, as in Figure 3.1. Here the *ID* attribute uniquely identifies the *Name* one, because by knowing the employee's ID we can tell what the employee's name is. Therefore, $ID \rightarrow Name$. More specifically, since also the employee's surname and gender are uniquely identified by the ID, we can write $ID \rightarrow Name, Surname, Gender$. Another example is provided in Table 3.1, in which $Temperature, pH, Season \rightarrow Ideal$.

Functional dependencies are a very well-known concept for data scientists, especially for those who work on relational models, and further details on

Orange Plantation			
Temperature	pH	Season	Ideal
28	7	Autumn	Y
20	7	Autumn	N
28	7	Winter	N
29	7.5	Autumn	Y
27	6.5	Winter	N
27	7.5	Summer	N
20	6.5	Spring	N
28	7	Summer	N
27	6.5	Autumn	Y

Table 3.1: 'Orange Plantation' table. It shows whether or not ambient temperature (°C), soil pH and planting season represent ideal conditions for planting oranges.

them can be found in [2]. However, the constraints imposed by functional dependencies are often too strict for real-world datasets (they must indeed hold for all the tuples of a table), so in the past few years generalizations of FDs have been proposed and started to be considered in their place. **Relaxed Functional Dependencies (RFDs)** can indeed be simply defined as functional dependencies where some constraints are deleted (relaxed). The authors of [11] distinguished 35 different categories of RFDs, but the ones relevant for our research are the following:

- **Approximate Functional Dependencies (AFDs)**:

  AFDs are FDs holding on almost every tuple. [11, p. 151]

  In order to quantify how an AFD 'almost' holds, several measures have been proposed, including the so-called *g3*, defined as "the (normalized) minimum number of tuples that need to be removed from a relation instance in order for an FD to hold" [11, p. 151], whereas 'relation instance' is simply a synonym of 'relation'. The g3 measure is therefore an index whose value ranges between 0 and 1, indicating the percentage of tuples of a table to be removed in order for an FD to hold (0 = none, 1 = all). An example of AFD in Table 3.1 is:

  $$Temperature, pH \rightarrow Ideal$$

  because almost all the values of *Temperature* and *pH* determine the *Ideal* value, but this is not true for:

  $$Temperature =\text{`` } 28 \text{ ''}, pH =\text{`` } 7 \text{ ''} \rightarrow Ideal$$

since in most of the cases (two out of three) when *Temperature* =‘ 28 ’ and *pH* =‘ 7 ’ then *Ideal* =‘ Y ’, but in one case *Ideal* =‘ N ’.

- **Conditional Functional Dependencies (CFDs)**:

  > [CFDs] *use conditions to specify the subset of tuples on which a dependency holds.* [11, p. 152]

  This type of dependencies allows to catch particular and concrete patterns in the dataset, in fact they make possible to analyze precise values of the tuples and be more specific. An example of CFD, related to Table 3.1, is the following:

  $$Temperature =‘ 28 ’, pH =‘ 7 ’, Season \rightarrow Ideal$$

  meaning that, for tuples in which *Temperature* =‘ 28 ’ and *pH* =‘ 7 ’, the *Season* parameter functionally determines the *Ideal* one. Another example could be:

  $$Season =‘ \text{Summer} ’ \rightarrow Ideal =‘ N ’$$

  interpretable as: ‘when the attribute *Season* has value Summer, the attribute value of *Ideal* is N’.

- **Approximate Conditional Functional Dependencies (ACFDs)**: FDs obtained by combining the two kinds of relaxed dependencies discussed above. Unifying the two relaxation criteria makes it possible to detect specific and not exact rules, which can highlight anomalies or unexpected patterns in the database, allowing to recognize cases where a value of a certain attribute *frequently determines* the value of another one. An example of ACFD in Table 3.1 is:

  $$Season =‘ \text{Autumn} ’ \rightarrow Ideal =‘ Y ’$$

  which can be read as: ‘when attribute *Season* has value Autumn, the attribute value of *Ideal* is Y if we delete a maximum number of tuples N from the dataset’. The rule indeed holds for almost all the tuples of the table, apart for the one in which *Temperature* =‘ 20 ’, *pH* =‘ 7 ’, *Season* =‘ Autumn ’ and *Ideal* =‘ N ’. It is worth to specify that, even though the notation is the same used for CFDs, the relaxation on the number of tuples is implicit in the classification of a rule as an ACFD.

3.6 Evaluation Metrics

The aim of this section is to introduce some evaluation metrics for functional dependencies used by one of the adopted tools, FAIR-DB, described later in Section 5.2.

- **Support**:

$$\text{Support}(X \to Y) = \text{supp}(X, Y) = \frac{\#(X, Y)}{\#tuples}$$

  where $\#(X, Y)$ is the amount of times the (sets of) attributes X and Y appear together in the dataset and $\#tuples$ is the total amount of tuples in the table. The support represents the percentage of records in the dataset that verify the dependency $X \to Y$, and it is therefore an index whose value ranges between 0 and 1.

- **Confidence**:

$$\text{Confidence}(X \to Y) = \text{conf}(X, Y) = \frac{\text{supp}(X, Y)}{\text{supp}(X)}$$

  where $\text{supp}(X)$ is the percentage of tuples in the dataset containing the (set of) attributes X (antecedent, or LHS, of the rule). The confidence shows how frequently the dependency $X \to Y$ is verified, knowing that the antecedent X is verified, and it is therefore an index whose value ranges between 0 and 1. A confidence equal to 1 means that only the valid and exact rules (non-relaxed FDs) will be selected, while decreasing its value implies relaxing the constraint on the number of tuples to be considered. In this context, it can be seen as an analogous metric to the g3 one, mentioned in Section 3.5.

- **Difference**:

$$\text{Difference}(X \to Y) = \text{diff}(X, Y) = \text{conf}(X, Y) - \text{conf}(X \setminus X_p, Y)$$

  where X_p is the subset of protected attributes of the antecedent X of the dependency $X \to Y$. The difference is basically a subtraction between the confidence value of a dependency and the confidence value calculated on the same dependency but excluding all the protected attributes from the antecedent of the rule. Being a subtraction of indexes of value between 0 and 1, and given $\text{conf}(X, Y) \geq \text{conf}(X \setminus X_p, Y)$, the difference is also an index whose value ranges between 0 and 1. It indicates how much a dependency is 'unethical' (the higher the value,

the more unfair is the dependency), and it gives an idea on the impact of the protected attributes on Y. Last but not least, it is important to point out that the difference is a novel metric, firstly introduced in [5] and specifically designed with the aim of measuring the 'ethical' level of a dependency.

3.7 Statistical Concepts

The aim of this section is to introduce some statistical concepts useful for fully understanding the behavior of one of the adopted tools, Ranking Facts, described later in Section 5.3.

The first required notion is the one of a **hypothesis test**, which in statistics is a way to test the obtained result of a survey or experiment on a sample, in order to check if the result is meaningful and extendable to the whole population or if it has happened by chance. In this context, two interpretations are proposed: the first is known as *null hypothesis* (symbolized as H_0), which is the idea that there is no relationship in the population and that the relationship in the sample is caused by errors (informally, this is the 'occurred by chance' interpretation); the second is called *alternative hypothesis* (whose symbol is H_1) and it is the idea that the relationship in the sample reflects an existing relationship in the population.

Generally, the rationale behind a hypothesis test is:

1. Assume the null hypothesis true.

2. Determine how likely the sample relationship would be if the null hypothesis were true.

3. If the sample relationship would be extremely unlikely, then *reject* the null hypothesis in favor of the alternative hypothesis. If it would not be extremely unlikely, then *retain* the null hypothesis.

A crucial step in hypothesis testing is to find the likelihood of the sample result if the null hypothesis were true. This probability is called ***p*-value**. Low p-value means that the sample result would be unlikely if H_0 were true and leads to the rejection of the null hypothesis, while high p-value means that the sample result would be likely if H_0 were true and leads to the retention of the null hypothesis. To quantify how low the p-value must be in order to consider the result unlikely enough to reject the null hypothesis, a parameter known as *significance level* α is used, and its value is usually set to 0.05 (5%). The significance level represents the probability of making the

mistake of rejecting the null hypothesis when in fact it is true (*type I error*): if p-value $> \alpha$ we accept the null hypothesis and the result is considered not statistically significant, otherwise we reject the null hypothesis and the result is said to be *statistically significant*.

A particular type of hypothesis test is the **z-test**, used when data are approximately normally distributed (i.e. the plotted data have the shape of a bell curve on the graph). In order for a z-test to be used, data points should also be independent of each other and the sample size should be greater than 30. The z-test relevant to our research is the *two sample z-test*, which allows to compare two proportions to check if they are the same (H_0) or not (H_1). The reference formula is:

$$z = \frac{p_1 - p_2}{\sqrt{(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})}} = \frac{p_1 - p_2}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

where n_1 and n_2 are the sizes of the samples, p_1 and p_2 are the proportions of the samples, p is the overall sample proportion (total number of 'positive' results over total number of people) and σ_1^2 and σ_2^2 represent the variances of the two populations. For the sake of completeness, we define the *variance* as the measure of how far each value in the dataset is from the average value (i.e. the mean, as defined in Section 5.1):

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}$$

where n is the size of the population, x_i is the i-th value and μ represents the mean. In the z-score formula above, for the sake of simplicity, σ_1^2 and σ_2^2 are approximated by the variance of the Bernoulli distribution (i.e. the probability distribution for a random experiment with only two possible outcomes). To better clarify, we will now consider the example of testing two different COVID-19 vaccines: the first works on 248 people out of a sample of 496, while the second works on 23 people in a sample of 100. To check if the vaccines are comparable we firstly calculate the proportions:

$$p_1 = 248/496 = 0.5$$

$$p_2 = 23/100 = 0.23$$

The overall sample proportion is:

$$p = \frac{248 + 23}{496 + 100} \simeq 0.45$$

And finally the z-score is:

$$z \simeq \frac{0.5 - 0.23}{\sqrt{0.45(1 - 0.45)(\frac{1}{496} + \frac{1}{100})}} \simeq 4.95$$

To find out if the obtained result should lead us to accept or reject the null hypothesis, we can look at the known values of z-score, related to the most commonly used α values. For example, the z-score related to $\alpha = 0.05$ is 1.96, meaning that 95% of the area under the normal curve lies within the range $[-1.96, +1, 96]$, and since $4.95 > 1.96$ we can reject the null hypothesis (for z-scores higher than 1.96 with $\alpha = 0.05$, p-value < 0.05).

Chapter 4

Sociological Research

The aim of this chapter is to provide the reader with a sociological background by reporting information about gender gap in the society, overall at first and then with a focus on the U.S., which may be useful to understand and better contextualize the reasons behind the presence of possible preexisting biases and unequal gender representation in the results of our analysis.

We will start by providing information about the *Global Gender Gap Index*, a powerful tool to measure different facets of the gender gap worldwide, and then we will focus on the U.S. by reporting some considerations about *gender discrimination in the workplace*, taken from the literature in the field. Finally, some *data and statistics from the U.S. Department of Labor* will be provided, with a brief comment on the displayed graphs.

4.1 The Global Gender Gap Index

A useful instrument to get a sociological overview on gender pay gap is provided by the World Economic Forum (WEF), "an independent international organization committed to improving the state of the world by engaging business, political, academic and other leaders of society to shape global, regional and industry agendas" [24]. The World Economic Forum periodically releases a document – The Global Gender Gap Report – which contains information about gender-based inequalities in the various countries of the world, and provides a ranking based on a cumulative measure called **Global Gender Gap Index**, defined as:

> *A framework for capturing the magnitude of gender-based disparities and tracking their progress over time. The Index benchmarks national gender gaps on economic, education, health and political*

criteria, and provides country rankings that allow for effective comparisons across regions and income groups. [44, p. 3]

The Index is based on three underlying criteria:

- It focuses on measuring *gaps rather than levels*, because in this way the Index is disassociated from countries' levels of development, which would heavily impact on the results. For example, rich countries are able to offer more education and health opportunities to all members of society, but this is quite independent of the gender-related gaps that may exist within those higher levels of health or education [44].
- It captures *gaps in outcome variables rather than gaps in input variables* (such as indicators related to country-specific policies, rights, or culture), in order to provide objective results based on fundamental measures related to basic human rights. For this reason, the Index relies on four categories (subindexes): Economic Participation and Opportunity, Educational Attainment, Health and Survival, Political Empowerment.
- It ranks countries according to *gender equality rather than women's empowerment*, because the focus is on the variation of the gap in the chosen indicators throughout the years. Therefore, the Index rewards countries that reach the point where outcomes for women equal those for men, but it neither rewards nor penalizes cases in which women are outperforming men.

The Global Gender Gap Report 2017 [44], besides the mere ranking, provides some sociological interpretations of the results, based on the subindexes outcomes and on the parameters of which they are composed, such as the employee educational attainment by level, field of study, and gender, or the share and evolution of female hires in various industries. Although the majority of considerations are generic in nature, and not specific for the U.S. society, they are worth to be mentioned, since most of them could be reflected, on a small scale, in our case studies.

- There is a current stagnation of progress towards closing the economic gender gap, for several reasons:
  - The global labor force participation has been in decline globally for both men and women, but this decline has been particularly accentuated for women.
  - Earned incomes of both men and women have been increasing, but this upward trend has been steeper for men than for women.

  - Women's share among senior positions both in the public sector and in business is not trending towards equal representation.

- There is an under-use of the ever-increasing numbers of educated women because of discrepancies in caregiving and unpaid work, institutional and policy inertia, outdated organizational structures, and discrimination, but also skill differentials in the types of degrees women and men seek out in their education.

- In many countries, a variety of social circumstances limit women's access to technology and therefore their ability to gain proficiency in its use. When women do have the relevant mathematical and technological skills, unconscious biases can influence their peers' recognition of their capabilities.

- There exist imbalances in the specific fields of study in which men and women tend to specialize. Women are underrepresented in the engineering, manufacturing, and construction as well as information, communication, and technology fields.

- There is a tendency towards lower pay for occupations that have historically developed as predominantly female. When women enter a profession in large numbers, the pay-related benefits of participating in the profession depreciate.

- The female leadership representation remains below 50% in all industries, and every industry exhibits a leadership gender gap.

- Unconscious biases and systemic efforts focused on driving change at the industry or country level through public-private collaboration remain scarce.

Appendix A shows the country profile of the United States in the report. As we can see, the country was ranked 49th out of 144, with an overall score of 0.718 (in a score system in which 1 means gender parity), and Political Empowerment is the subindex in which the U.S. performed worst. For the purpose of our research, it is particularly useful to look at the *Economic participation and opportunity* section of the country score card, in which we can notice that men are more participatory in the labor force than women, and women tend to earn less then men for similar jobs. Furthermore, the estimated earned income for women is significantly lower than the one of men, and the last two items show that women are underrepresented in managerial positions and higher-paying jobs, and are overrepresented in professional and

technical works. From the *Workforce Participation* section of the selected contextual data we can see that women are more likely than men to be employed part-time, and also not to be paid for their work.

Other considerations about the gender gap, as depicted in The Global Gender Gap Report 2017, are done by the authors of [29]. Even if the most relevant for us are the ones concerning the Economic Participation and Opportunity subindex, we believe that for a broader understanding it is worth to briefly report also the main points related to the other subindexes, especially because they have an impact on the condition of women which can also affect the workplace:

- **Economic Participation and Opportunity**: women seem to be more likely than men to be living at or below poverty, mainly because of the following reasons:
  - Many women remain economically dependent on men.
  - Women are more likely than men to be unemployed or to work in positions in which they do not get paid.
  - Women are more likely than men to be concentrated in industries and occupations with low wages, long hours, and no social protections, and less likely than men to hold management positions.
  - Women in general earn only 82% compared to white men, and the gender wage gap becomes further complicated when race/ethnicity is taken into account.
- **Educational Attainment**: women are concentrated in traditionally female and lower-paying CTE (Career and Technical Education) programs in both secondary and postsecondary educational settings, and are still underrepresented in Science, Technology, Engineering, and Math (STEM) programs. Gender stereotypes and bias in education and the potentially hostile climate of academic departments continue to deter women from these lucrative career opportunities.
- **Health and Survival**: women are at a disadvantage compared to men, mainly because of the following reasons:
  - Poor access to information, early marriage, lack of decision-making power continue to increase women's exposure to sexually transmitted diseases, unwanted pregnancies and the risk of unsafe abortions.

  - Women are constantly bombarded with media advertisements that sexualize their bodies. The influence of media, television, movies, etc. has led to increased prevalence of body dissatisfaction and eating disorders globally.

  - Violence towards women continues to impact women's health worldwide, and makes it difficult for women to pursue educational opportunities or to perform their jobs.

- **Political Empowerment**: women hold a minority of political and institutional decision-making positions. Gender norms and prejudices work to both reduce the number of female candidates and contribute to the obstacles faced by women in elections.

Finally, we believe it is worth to mention that the Global Gender Gap Index is not the only tool to evaluate gender equality among countries, and another example is represented by the *Historical Gender Equality Index (HGEI)*, introduced in [19]. As the name itself suggests, HGEI is based on some historical measures, since gender inequality is strictly related to the human history, and it has the aim of providing a global overview of gender equality in the long run, as well as to give an indication of gender disparities in well-being outcomes that result from institutional, cultural, and social influences. Similarly to the Global Gender Gap Index, HGEI is also based on a few requirements – coverage (of gender equality dimensions), availability of data for many countries, simplicity in calculation and understanding, possibility of comparisons between countries but also over time – and it is a composite index made of four dimensions – health, autonomy within the household, political power, socioeconomic status – each of which composed by some indicators. As the Global Gender Gap Index, also HGEI is based on ratios rather than levels, in order to evaluate the position of women relative to men in each society rather than the actual levels of resources and opportunities available to women, and therefore it does not capture how women are doing in absolute terms, and it cannot show if there are cases where women are outperforming men. HGEI revealed that most countries of the world made progress toward gender equality over the past fifty years, but there is little convergence between them, and the authors recommend that future research should also pay attention to the dimensions in which gender inequality occurs, because behind a composite index can lie great variation in the underlying indicators.

4.2 Gender Discrimination in the Workplace

The Global Gender Gap Index is a powerful indicator, which provides us with an overview on the main fields in which women experience inequalities, particularly on a large scale, and with some sociological motivations for these disparities. We will now focus on some literature regarding the U.S. society and more specific on the labor market.

Tilcsik in [47] brings into play the theory of **statistical discrimination**, which rather than merely explaining discrimination, helps rationalize and justify discriminatory decisions. As reported in the article:

> *This theory* [statistical discrimination] *posits that employers have imperfect information about the future productivity of job candidates, which gives them an incentive to use easily observable ascriptive characteristics, such as race or gender, to infer the expected productivity of applicants.* [...] *In this model, discrimination does not arise from animus or antipathy toward members of a group; rather, it is portrayed as a rational solution to an information problem.* [47, p. 94]

Statistical discrimination is in contradiction with the other dominant economic perspective on discrimination, known as **taste-based model**:

> *Unlike statistical discrimination theory, which taps into the culturally valued discourse of instrumental rationality and frames discrimination as a logical solution to an information problem, the taste-based model is about negative attitudes, such as overt racial prejudice and sexism, that tend to be publicly disavowed and are often perceived as socially unacceptable.* [47, p. 95]

Tilcsik focused on some sociological perspectives in support of the statistical discrimination theory, explaining that status beliefs and stereotypes shape how employers evaluate workers and how they distribute rewards among them. Indeed, when employers are required to evaluate groups of candidates for a job position, their perceptions of the differences among the groups reflect cultural beliefs that are often inaccurate and resistant to change, even in the face of disconfirming evidence, and therefore it is not a matter of intentional discrimination but of societal bias.

Statistical discrimination is likely to have some resonance and normative acceptability for three main reasons [47]:

- It is a rational, profit-maximizing, incentive-driven decision, and represents 'the optimal solution to an information extraction problem'.

- Some economists characterize it as fair and morally defensible, by implying that categorical differences are rooted in statistical considerations that rational employers consider, and by suggesting that statistical discrimination is fair and neutral because it treats people with the same expected productivity identically.
- Economists often emphasize that it is ubiquitous and practically inevitable in many domains of life.

Thus, the use of stereotypes is depicted as cognitively and economically useful as well as consistent with social norms. Exposure to the idea of statistical discrimination strengthen people's belief in the validity, usefulness, and acceptability of relying on stereotypes and hence increase their likelihood of engaging in discrimination because of ascriptive group characteristics. When employers feel confident that their decisions are impartial, rational, and ethically defensible, they feel more justified in relying on stereotypes and exert less effort to suppress their biases. Statistical discrimination may also become self-fulfilling if it leads members of negatively stereotyped groups to believe that investing in their skills will not be fully rewarded.

Tilcsik in [47] also conducted a a survey experiment consisting in a hiring simulation, in which participants (who all had managerial experience) were randomly assigned to one of four conditions:

1. Exposure to statistical discrimination theory (treatment).
2. No exposure to any theory of discrimination (non-treatment).
3. Exposure to the taste-based model (placebo).
4. Exposure to statistical discrimination theory and a critical commentary (treatment variant).

Participants exposed to the theory (without a critical commentary) perceived stereotyping as more acceptable and stereotypes as more accurate than did participants in the other groups, and selected fewer women for their teams. Group representation was also impacting on decisions, since female participants and those who did not identify as either male or female were, on average, less convinced of the acceptability and accuracy of stereotypes than were male respondents.

Beggs in [7] instead shifts the focus on the impact of the **institutional environment** on gender and race inequalities in the labor market. According to the theoretical background he depicts, "organizations compete not just for resources and customers, but for political power and institutional legitimacy,

for social as well as economic fitness" [7, p. 613], from [20, p. 150]. Therefore, an important factor to consider is the force of public opinion: when new definitions or practices become legitimated and accepted, organizations are under considerable pressure to incorporate them.

Beggs decided to analyze data from the 1980 U.S. Census in order to prove the two hypotheses reported below [7]:

1. Within industries, the higher the proportion of workers employed in states with high support for equality, the lower the levels of race and gender inequality in jobs and earnings.

2. The higher the proportion of federal public sector employees in an industry, the lower the levels of race and gender inequality in jobs and earnings.

For the analysis, he decided to adopt some measures, each of which composed by several indicators: local institutional environment and national institutional environment (independent variables), quality of employment and earnings inequality (dependent variables), industrial structure, human capital inequality, and employment inequality (control variables).

According to his results, the institutional environment impacts both quality of employment and earnings. For what concerns the former, greater support for equality in the local institutional environment, as well as greater federal public sector employment in an industry, is associated with lower levels of inequality among minorities (race/gender groups). As for the latter, the greater the support for equality in the local institutional environment, the better the earnings position of each minority relative to white men, and the level of federal public sector employment in an industry is positively associated with the earnings position of all minorities.

Finally, an interesting point of view is provided in [22], a report published by the Economic Policy Institute. Folbre highlights that women in the U.S. still tend to earn 20% less per hour than men, and points out that empirical research on the causes of the persistent earnings gap often takes the form of statistical models that control for as many variables as possible (such as race, education, labor force experience), but control variables cannot be explained purely as the result of individual choices. Rather, they reflect structural inequalities related to **unequal bargaining power**.

> *Bargaining often characterizes situations where two parties, whether individuals or groups, see potential gains from cooperation but disagree over how those gains should be shared. Both parties can potentially benefit from coming to an agreement, and their*

share is likely to be strongly affected by their fallback position, or next-best option. [22, p. 9]

We can summarize the main points of Folbre's research (that is, the main causes of gender gap) as follows:

- **Stereotypes**: in the U.S. women who move into better paying but stereotypically masculine occupations often face sexual harassment and disapproval. When wives earn more than husbands, both spouses slightly tilt their reported earnings to conform to gender stereotypes, overstating the relative size of husbands' earning.

- **Pay penalties**: many women self-select into traditionally female occupations because they consider these more compatible with the demands of family care. But while they may be aware that these jobs pay less than traditionally male jobs, they do not choose the size of the resulting pay penalties: in the U.S. the percentage of women in an occupation is inversely related to its average pay, even controlling for human capital characteristics.

- **Mothers discrimination**: many women experience wage penalties because of becoming mothers. Discrimination against mothers is based on rather subtle cues, such as participation in a parent-teacher organization listed on a job resume, because employers may assume that mothers of young children face other demands on their time that lower their performance in paid employment.

- **Mobility**: women's labor supply is less elastic than men's because women's mobility between jobs is limited by obligations of family care. Employers can easily take advantage of this difference, paying women less than men not because they prefer hiring men but because they recognize that women are more likely to accept lower wages.

- **Occupational segregation**: high levels of occupational segregation are still the largest immediate cause of gender inequality in earnings. Efforts to improve the relative pay of primarily female jobs met criticism based on the assumption that occupational pay was largely determined by productivity. Finally, the supply of women's labor to the market was treated merely as the result of individual choices, with little attention to the constraints imposed by a traditionally male-oriented organization of work, school, housework, and childcare.

- **Information asymmetries**: employers are often able to use their control over information to lower women's wages relative to men's, because not all the workers are covered by the Fair Labor Standards Act (which guarantees the right of workers to discuss their salaries), and others remain unaware of their rights. Furthermore, employers in most states have the right to ask job applicants what they earned in their previous jobs and to adjust wage offers accordingly. For this reason, women's entrance into previously male-dominated occupations tends to lower the average occupational wage, and this 'devaluation' is at least partially driven by the fact that women start out in lower-paying (often part-time) occupations, which lowers their bargaining power.

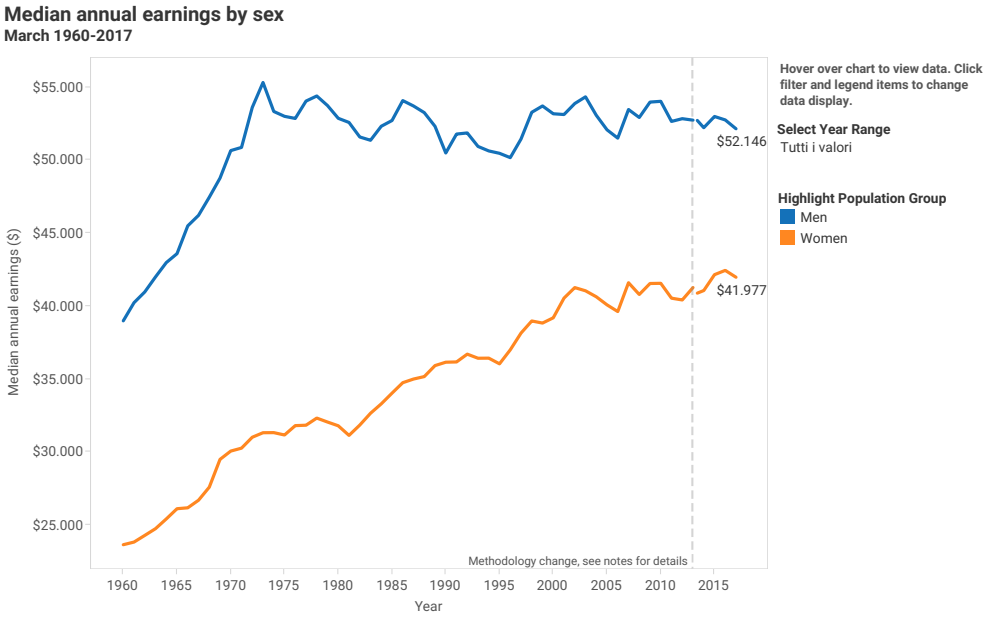
- **'Equal value' principle**: whatever a worker is paid represents his or her value added – therefore, men earn more than women because they contribute more to society. Unfortunately, this principle does not take into account that many female-dominated services have a public good dimension: their social value exceeds their private value. The work of caring for others creates value that is difficult to capture through the market because it often involves emotional engagement, teamwork, and person-specific skills. Women's tendency to devote more hours to unpaid care work than do men is interpreted because of feminine preferences rather than as the result of institutional pressure to ensure a generous supply of female effort to activities such as family care that cannot be rewarded by market forces.

- **Paid work/family work constraints**: responsibility for the care of family, friends, and neighbors weighs more heavily on women than men, not because women necessarily prefer this arrangement but because men often have sufficient bargaining power to minimize demands on their time. Furthermore, women labeled as uncaring are typically stigmatized, and employers use this social norm to justify lower pay offers to women.

4.3 Data & Statistics (U.S. Department of Labor)

The last step of our sociological research consists in the reporting of some interesting data and statistics[1] taken from the U.S. Department of Labor, and more specifically from Women's Bureau, an agency within the Department with the aim of developing policies and standards and conducting inquiries

[1] Available at: `https://www.dol.gov/agencies/wb/data`.



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

Figure 4.1: Median annual earnings by sex (1960–2017). U.S. Department of Labor. Source: `https://www.dol.gov/agencies/wb/data`.

to safeguard the interests of working women, to advocate for their equality and economic security for themselves and their families, and to promote quality work environments.

First of all, Figure 4.1 shows that, since 1960, both male and female earnings increased by about \$15K, but, because of the large initial gap, the actual average incomes still differ by about \$10K in favor of men. Furthermore, the graph does not take into account part-time workers, the introduction of which would lead to an accentuation of the gap between men and women, since, as Figure 4.2 displays, most part-time workers are women (over 60% of the total number of part-time employees).

Figure 4.3 shows the most common occupations for women. As we can see, in accordance with the other sociological sources, the podium is constituted by nurses, teachers, and secretaries. The representation problem is even exacerbated when looking at Figure 4.4, which is interesting for us because it displays the top-10 occupations employing the largest number of women for the same year without excluding part-time employees. Even

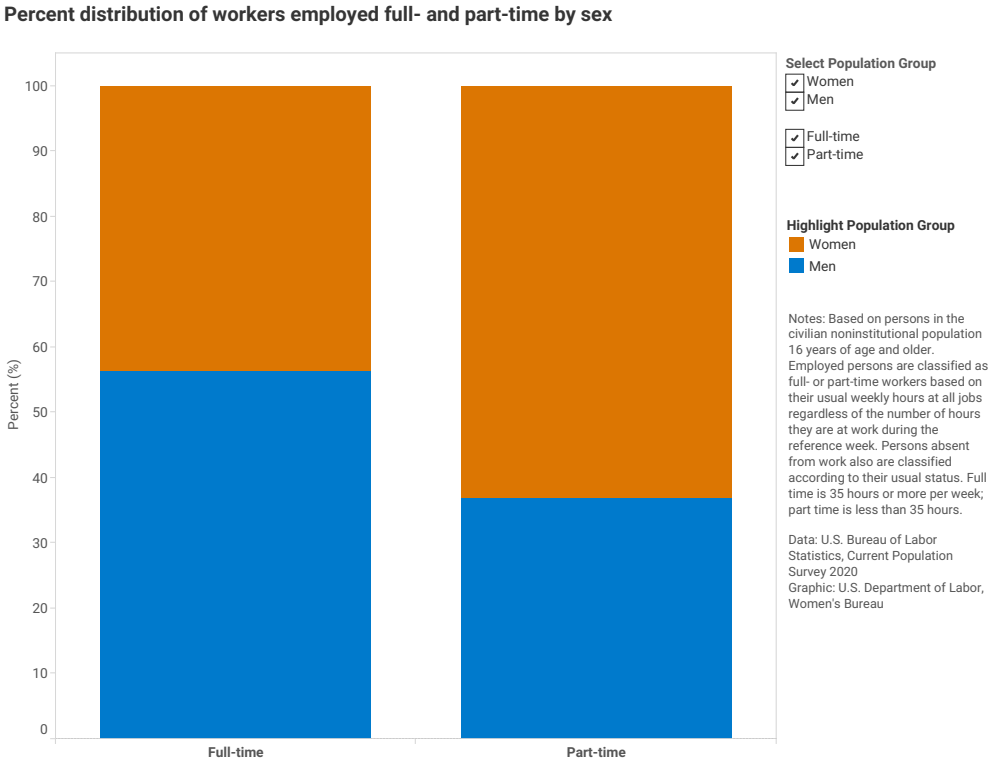


Figure 4.2: Percent distribution of workers employed full- and part-time by sex (2020). U.S. Department of Labor. Source: `https://www.dol.gov/agencies/wb/data`.

though data are a bit more aggregated (for example, the teachers category presumably includes not only elementary and middle school teachers), the numbers, in comparison with Figure 4.3, dramatically increase, emphasizing again the huge percentage of women working part-time. This problem is particularly accentuated for some job titles, typically not very profitable (such as cashiers and waitresses), which overcome more advantageous positions (like managers) in the list. Figure 4.5 instead highlights again how women are underrepresented in STEM disciplines, especially in computer occupations and engineering. Although the trend is generally increasing, the rate of increase is very low (+4% in 30 years, from 1990 to today), and it is particularly dramatic to observe the decrease in the percentage of female employees working in computer occupations, which had a peak – never reached again – back in 1990.

Finally, Figure 4.6 displays the ordered list of occupations with the largest gender earnings gap. Most of the professions are traditionally lower-paying jobs, and no STEM or managerial disciplines appear to be listed. As a consequence, considering the lower percentage of women employed in STEM

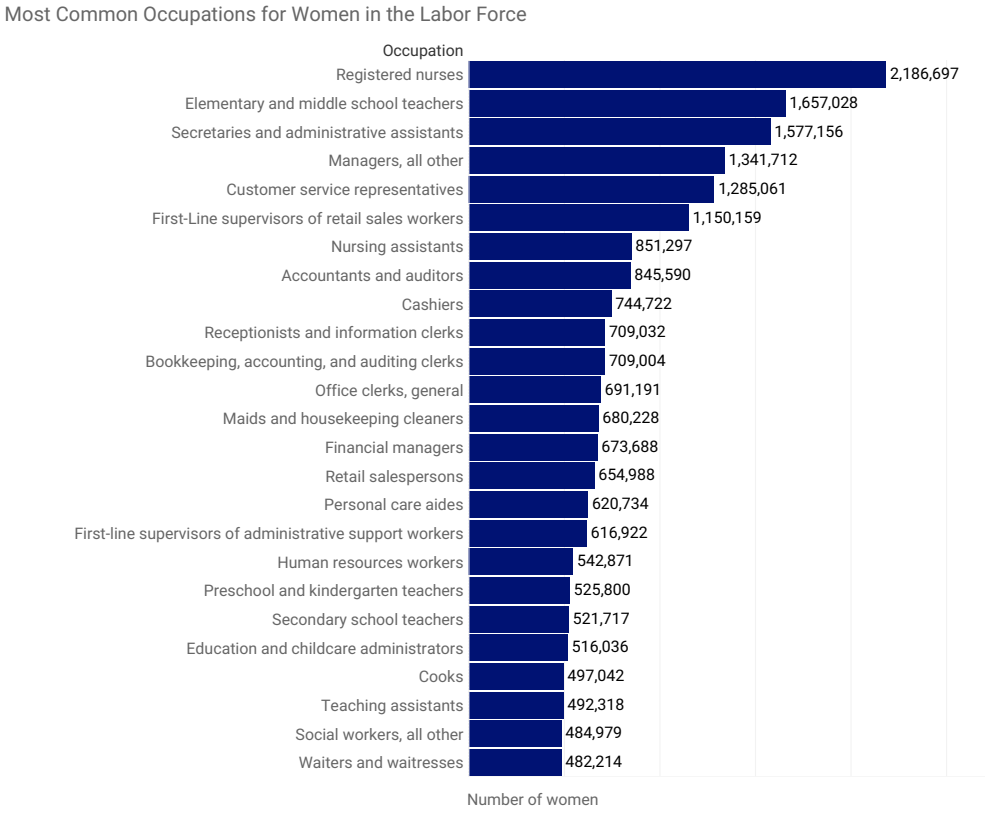


Figure 4.3: Most common occupations for women in the labor force (2019). U.S. Department of Labor. Source: `https://www.dol.gov/agencies/wb/data`.

and managerial positions, and considering that in general most part-time workers are employed in lower-paying jobs, we can put emphasis on the percentage of women being disadvantaged in comparison with men, and on the still present significance of the gender gap.

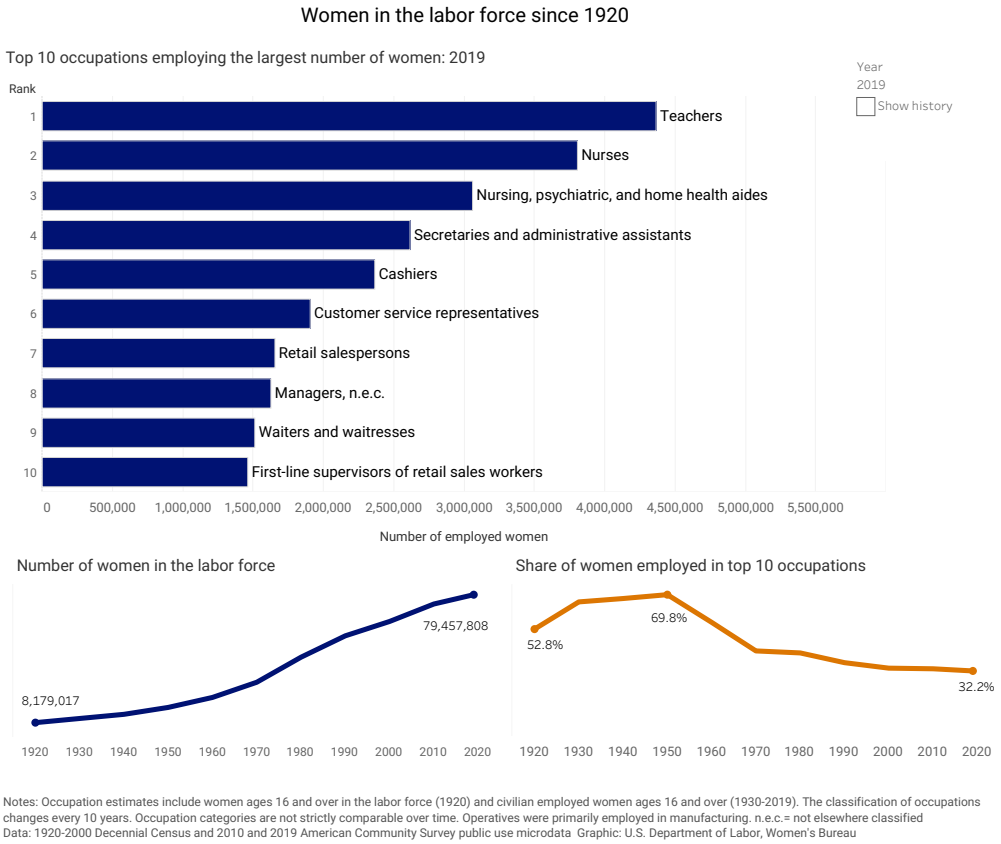


Figure 4.4: Top-10 occupations employing the largest number of women (2019). U.S. Department of Labor. Source: `https://www.dol.gov/agencies/wb/data`.

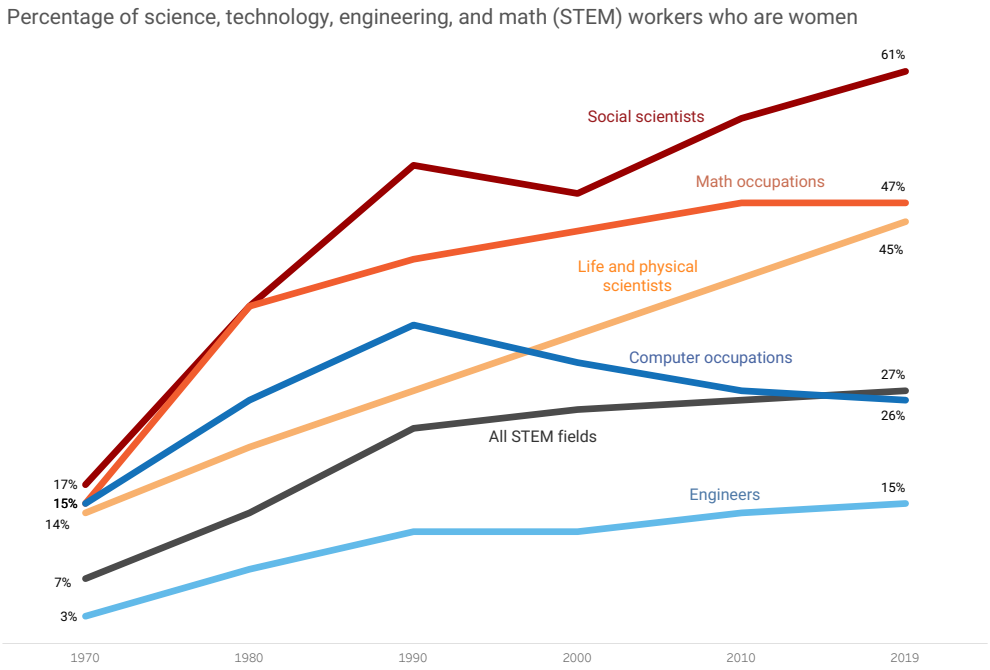


Figure 4.5: Percentage of Science, Technology, Engineering, and Math (STEM) workers who are women (1970–2019).
U.S. Department of Labor. Source: `https://www.dol.gov/agencies/wb/data`.

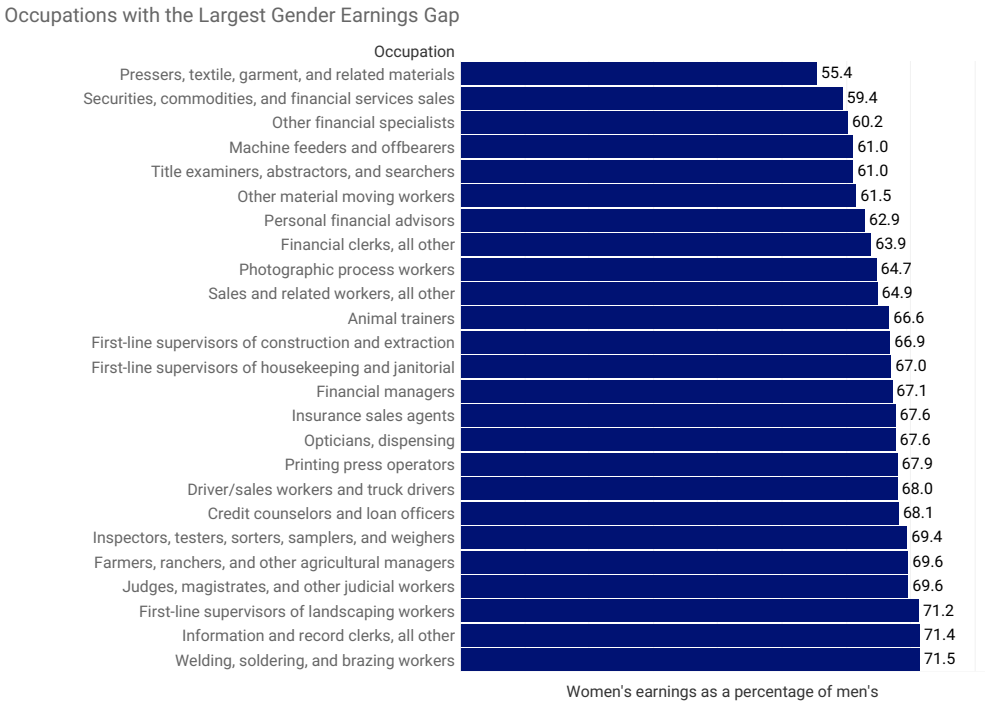


Figure 4.6: Occupations with the largest gender earnings gap (2019). U.S. Department of Labor. Source: `https://www.dol.gov/agencies/wb/data`.

Chapter 5

Techniques

The aim of this chapter is to provide the reader with an overview on the tools adopted in our experiments, recalling the technical concepts addressed in Chapter 3.

We will describe the functioning behind:

- *The 'Glassdoor Method'*, a framework for evaluating gender pay gap which relies on linear regression.
- *FAIR-DB*, an algorithm based on functional dependencies and the related evaluation metrics.
- *Ranking Facts*, an application built on the idea of ranking which makes use, among other things, of some statistical concepts previously introduced.

5.1 The 'Glassdoor Method'

After having explored the technical basics in Chapter 3, we will now make an overview of the tools used for this research. The first one is a technical guide to analyze gender pay gap in a company, provided by **Glassdoor** in [12].

Glassdoor is a website in which employees and ex-employees of companies anonymously review enterprises and their superiors, with the overall aim of providing insights about jobs and companies and helping people find the most suitable working position for them. The society was founded in 2007 in the U.S. and the website was made available in 2008; since then, the company has grown to become the worldwide leader in the sector.

As specified in the introduction of the guide [12, p. 2], according to a 2016 Glassdoor survey, 67% of the U.S. employees would not apply for jobs at employers where they believe a gender pay gap exists. The purpose of the

report is therefore to help HR practitioners in analyzing the internal gender pay gap of their companies, by providing them specific technical knowledge.

First of all, by 'gender pay gap' Chamberlain means:

> *The difference between average pay for men and women, both before and after we've accounted for differences among workers in education, experience, job roles, employee performance and other factors aside from gender that affect pay.* [12, p. 3]

Two measures are proposed in the report, respectively referred as 'unadjusted' and 'adjusted' pay gap:

- **'Unadjusted' pay gap**:

  > *Average pay for men as a group, compared to average pay for women as a group.* [12, p. 3]

  Therefore, the formula for estimating 'unadjusted' gender pay gap is:

  $$U = \frac{\text{avg}(BasePay)_m - \text{avg}(BasePay)_f}{\text{avg}(BasePay)_m}$$

  where $\text{avg}(BasePay)_m$ is the average pay (arithmetic mean of salaries) of male employees, while $\text{avg}(BasePay)_f$ is the average pay of female employees. For the sake of completeness, despite being a basic mathematical concept, we define the *mean* as the sum of a collections of numbers (in this case, salaries) divided by the count of numbers in the collection (the total amount of males or females). Taking n as total number of male employees:

  $$\text{avg}(BasePay)_m = \frac{\sum_{i=1}^{n} BasePay_i}{n}$$

  and the same holds for female employees.

- **'Adjusted' pay gap**: while the 'unadjusted' pay gap is basically a simple comparison of all women with all men, the 'adjusted' pay gap compares similarly situated male and female employees, in order to include in the calculation the numerous factors that affect the pay (e.g. job title, or educational level). The estimation of the 'adjusted' pay gap is based on linear regression, a concept presented in Section 3.4, and the reference formula is:

  $$y_i = \beta_1 Male_i + \beta_2 X_i + \epsilon_i$$

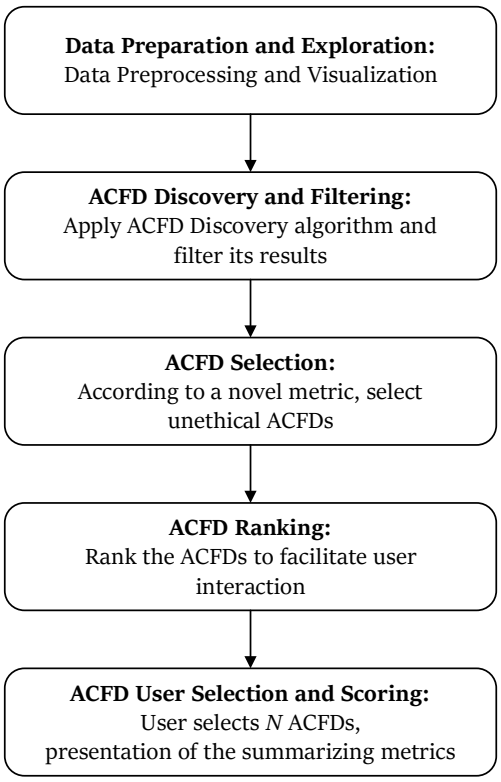


Figure 5.1: Steps of the FAIR-DB framework. Image based on the one shown in [5].

where y_i is the annual salary of worker i, $Male_i$ is a dummy indicator equal to 1 for males and 0 for females, and X_i is a collection of attributes of employees which may be relevant for the calculation (job title, educational level, etc.). The estimated coefficient β_1 represents the approximate pay advantage for men compared to women.

5.2 FAIR-DB

The second tool adopted for this research is called **FAIR-DB** (***F**unction**A**l Dependenc**I**es to discove**R** **D**ata **B**ias*), and it is a framework with the aim of discovering unfair behaviors in datasets, developed at Politecnico di Milano. As the name itself suggests, FAIR-DB is based on functional dependencies, and it falls within the category of preprocessing techniques since it works by finding conditions (constraints) already present in the data. The developers documented the functioning of FAIR-DB in [5], by providing an overview of the tool, together with a clarifying example.

Figure 5.1 shows the framework workflow, while a brief explanation of each phase, as documented in [5], is reported below.

- **Data preparation and exploration**: data are imported and data integration is eventually performed. Data cleaning, feature selection and discretization techniques are also applied in this phase, in order to deal with missing values, select the smallest set of attributes relevant for the analysis and transform data from numerical to nominal data type. Data are finally plotted in order to help the user in identifying groups in the dataset and eventually majority and minority classes.

- **ACFD Discovery and filtering**: the *ACFD Discovery* algorithm, presented in [41], is applied to extract approximate conditional functional dependencies from the dataset. The algorithm takes as input the dataset and three threshold parameters: *minimum support*, *minimum confidence* and *maximum antecedent size* of the ACFD sought. From the output, dependencies not involving at least one of the protected attributes and the target attribute (the one used as reference to search for discrimination, e.g. *Income*) are removed, as well as dependencies containing variables (in which one or more attributes are not assigned to a specific value, e.g. the attribute *Ideal* in *Temperature* =‘ 28 ’, *pH* =‘ 7 ’→ *Ideal*).

- **ACFD selection**: for each ACFD, some metrics are computed to capture the ‘ethical level’ of the dependency. In particular, the *difference* metric described in Section 3.6, as already mentioned, is a novel score introduced for this purpose, and a second measure called *p-Difference* is calculated for each protected attribute. The p-Difference indicates how much a dependency shows bias with respect to a specific protected attribute, and it is computed in the same way as the difference, but excluding the attribute from the antecedent of the rule. According to the values of the metrics, the most interesting ACFDs are selected.

- **ACFD ranking**: the ACFDs are ranked in descending order of importance according to *support*, *difference*, or *mean*. The support emphasize the *pervasiveness* of a rule, because it indicates the number of tuples involved by the dependency, so the higher the value, the more tuples are affected by the ACFD. The difference privileges the *unethical aspect* of a rule, because it highlights dependencies where the values of the protected attributes influence most their RHS. The mean is computed as mean of support and difference, and therefore it gives more impor-

tance to the rules with the *best trade-off* between pervasiveness and unethical perspective.

- **ACFD user selection and scoring**: the user selects N ACFDs perceived as the most problematic, and the system computes metrics (based on support, difference, and p-Difference of the selected rules) to summarize the level of unfairness of the dataset.

5.3 Ranking Facts

The third tool used for this research is called **Ranking Facts**, a Web-based application (of which there also exists a notebook version) developed by the team of *Data, Responsibly*[1]. Ranking Facts, as the name itself suggests, is a *ranking* tool: ranking is an action commonly performed by the vast majority of the algorithms we use every day: Google itself ranks the results of our searches and provides us with a list in descending order of relevance, and the same mechanism is used in various contexts of different nature, like dating or hiring applications. These specific scenarios are particularly relevant, because it is people who are ranked, and therefore discrimination against individuals or protected groups could arise, or the outcome could exhibit low diversity. Ranking Facts is based on the concept of *nutritional labels*, in analogy to the food industry, where simple, standard labels convey information about the ingredients and production processes. Similarly, in the tool nutritional labels are derived as part of the complex process that gave rise to the data or model they describe, embodying the paradigm of interpretability-by-design.

As documented in [50], Ranking Facts is a collection of visual widgets with the aim of providing to the user information about the ranking in terms of stability, fairness and diversity. A brief description of how they work, taken from [50], is reported below.

- **Recipe** and **Ingredients**: the former widget succinctly describes the ranking algorithm, by listing the attributes used for ranking together with their weights, as specified by the user; while the latter shows, in descending order of importance, the attributes that really affect the ranking.

- **Stability**: it explains whether the ranking methodology is robust on the specific dataset in use. An unstable ranking is one where slight changes to the data (e.g. due to uncertainty and noise), or to the

[1] Available at: `http://demo.dataresponsibly.com/rankingfacts`.

methodology (e.g. by slightly adjusting the weights of the attributes in the recipe) could lead to a significant change in the output.

- **Fairness**: it quantifies whether the ranked output exhibits statistical parity (group fairness) with respect to one or more protected attributes, such as gender or race of individuals. The notion of fairness is defined specifically for rankings and it can be computed comparing only binary categorical attributes (i.e. non-numerical attributes with just two possible values). The summary view of the widget presents the output of three fairness measures:

  - **FA*IR** [52]: ranking algorithm based on the assumption that on a ranking, the desired good for an individual is to appear in the result and to be ranked among the top-k positions. The outcome is therefore unfair if members of a protected group are systematically ranked lower than those of a privileged group, and a ranking algorithm discriminates unfairly if this ranking decision is based fully or partially on a protected feature.

    The *ranked group fairness* criterion used by the algorithm compares the number of protected elements in every prefix of the ranking (i.e. the top-i positions of the ranking, with $i \in [1, k]$) with the expected number of protected elements if they were picked at random using Bernoulli trials (independent 'coin tosses') with success probability p. The statistical test also includes a significance parameter α, corresponding to the probability of a type I error, which means rejecting a fair ranking. A clarifying example is provided in Table 5.1.

    The algorithm produces a top-k ranking that satisfies the ranking group fairness criterion mentioned above while maximizing *utility*, which means selecting the 'best' tuples, assigning them a score based on the relevant attributes used for the evaluation (e.g. picking the most qualified candidates for a job position by looking at their educational level).

  - **Proportion** [53]: this measure is based on the concept of z-test, as described in Section 3.7.

  - **Pairwise**: also known as *pairwise comparison*, it is a tool for prioritizing and ranking multiple options relative to each other. A matrix is generally used to compare each option in pairs and determine which is the preferred choice or has the highest level of importance based on defined criteria. At the end of the comparison

p \ k	1	2	3	4	5	6	7	8	9	10	11	12
0.1	0	0	0	0	0	0	0	0	0	0	0	0
0.2	0	0	0	0	0	0	0	0	0	0	1	1
0.3	0	0	0	0	0	0	1	1	1	1	1	2
0.4	0	0	0	0	1	1	1	1	2	2	2	3
0.5	0	0	0	1	1	1	2	2	3	3	3	4
0.6	0	0	1	1	2	2	3	3	4	4	5	5
0.7	0	1	1	2	2	3	3	4	5	5	6	6

Table 5.1: Minimum number of candidates in the protected group that must appear in the top-k positions to pass the ranked group fairness criterion with $\alpha = 0.1$. Considering for example $gender$ as a protected attribute with values ' M ' and ' F ', the minimum number of females (or eventually males) appearing in the top-5 with $p = 0.4$ is 1. Table based on the one shown in [52].

	Coffee	Wine	Tea	Beer	Sodas	Milk	Water
Coffee	1	9	3	1	$\frac{1}{2}$	1	$\frac{1}{2}$
Wine	$\frac{1}{9}$	1	$\frac{1}{3}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
Tea	$\frac{1}{3}$	3	1	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{4}$	$\frac{1}{5}$
Beer	1	9	4	1	$\frac{1}{2}$	1	1
Sodas	2	9	5	2	1	2	1
Milk	1	9	4	1	$\frac{1}{2}$	1	$\frac{1}{2}$
Water	2	9	5	1	1	2	1

Table 5.2: Drink consumption in the U.S. represented in a pairwise comparison matrix. Table based on the one shown in [42].

process, each option has a rank or relative rating as compared to the rest of the options. Table 5.2 provides an example: scores are assigned based on how strongly the consumption of a drink on the left dominates that of a drink at the top. For example, when coffee on the left is compared with wine at the top, since coffee appears to be extremely more consumed, 9 is entered in the first row and second column position. A score of $\frac{1}{9}$ is automatically entered in the second row and first column position. According to the matrix, the final ranking would be:
Sodas : 0.252, *Water* : 0.228, *Beer* : 0.164, *Milk* : 0.148, *Coffee* : 0.142, *Tea* : 0.046, *Wine* : 0.019.

All these measures are statistical tests, and whether a result is fair is

determined by the computed p-value.

- **Diversity**: since fairness is also related to representation, this widget shows diversity with respect to a set of demographic categories of individuals, or a set of categorical attributes of other kinds of items, by displaying the proportion of each category in the top-10 ranked list and overall (i.e. considering all the elements in the ranking).

Chapter 6

Experiments

The aim of this chapter is to describe the experiments conducted using the tools introduced in Chapter 5, in order to verify the presence (and the nature) of bias in our datasets and discuss their fairness, according to the concepts introduced in Chapter 2 and the sociological background explored in Chapter 4.

For each case study, we will provide a *dataset description*, in order to allow the reader to understand how the original dataset looks like, then we will discuss about our *data preprocessing* choices, and finally about how the tools performed on the preprocessed dataset, providing further details on the algorithms when needed. Lastly, we will question the results obtained by considering the impact of *other design choices*, checking the impact on the used tools when different decisions are made.

6.1 Case Study 1: Chicago

6.1.1 Dataset Description

The main purpose of this research is to combine the technological perspective with the sociological one, in order to analyze the strengths and weaknesses of the adopted tools in real-world scenarios. For this reason, we decided to use real-world datasets, containing information related to public employees of the U.S., and more specifically of public employees working in the cities of Chicago[1] and San Francisco[2]. It is worth to specify that these datasets exist because of the Freedom of Information Act (FOIA): a federal law constituting

[1] Available at: `https://www.chicago.gov/city/en/depts/dhr/dataset/current_employeenamessalariesandpositiontitles.html`.

[2] Available at: `https://www.kaggle.com/tomtillo/san-francisco-city-payrollsalary-data-20112019`.

Title 5 of the United States Code (5 U.S.C. §552), which claims that federal employee salaries must be public information under open government laws.

The **Chicago** dataset we considered includes 31,858 tuples and is made up of 8 attributes, briefly described as follows:

- *Name*: full name of the employee in the form of 'Surname, Name'.
- *Job Titles*: categorical variable representing the job title of the employee (e.g. POLICE OFFICER). There are 1089 distinct values.
- *Department*: categorical variable representing the job department where the employee works (e.g. POLICE). There are 36 distinct values.
- *Full or Part-Time*: binary categorical variable describing whether the employee is employed full-time (F) or part-time (P).
- *Salary or Hourly*: binary categorical variable describing whether the employee is paid on an hourly basis or a salary basis. Employees paid on an hourly basis are further defined by the number of hours they work in a week.
- *Typical Hours*: numerical variable describing the typical amount of work (in terms of number of hours per week) for employees paid on an hourly basis. For employees paid on a salary basis the attribute value is null.
- *Annual Salary*: numerical variable describing the annual salary rate. It only applies for employees whose pay frequency is Salary, while for those whose pay frequency is Hourly the attribute value is null.
- *Hourly Rate*: numerical variable describing hourly salary rates for employees whose pay frequency is Hourly. For employees whose pay frequency is Salary the attribute value is null.

6.1.2 Data Preprocessing

In order to simplify the subsequent bias analysis, we operated some **data transformation** processes on the attributes, choosing what we believe to be the most suitable names. For the Chicago dataset, we renamed *Job Titles* to *Job Title* and *Full or Part-Time* to *Status*. We also performed some **data aggregation**, estimating the *Annual Salary* of employees paid on an hourly basis by using the formula *Typical Hours* $\times$ *Hourly Rate* $\times$ 52, where 52 is a constant representing the number of weeks in a year.

Since our focus is gender pay gap but the original datasets do not contain a *Gender* attribute, we adopted a Python package called `gender-guesser`[3]. The aim of the package is to infer a person's gender from their first name, and the possible outcomes are: unknown (name not found), andy (androgynous), male, female, mostly_male, or mostly_female. The difference between andy and unknown is that the former is found to have the same probability to be male than to be female, while the latter means that the name was not found in the database. For each employee, we split the *Name* attribute to obtain their *First Name*, and then we inferred their gender by using the package. We obtained (out of the total of 31,858 tuples):

- unknown: 2,653 values.
- andy: 184 values.
- male: 20,562 values.
- female: 6,954 values.
- mostly_male: 775 values.
- mostly_female: 730 values.

In order to get coherent results in case of multiple experiments on the same dataset, we decided to remove the tuples related to unknown and androgynous names instead of randomly assign a gender to them (otherwise, we would have get different numbers at each execution of the preprocessing algorithm). Furthermore, we assumed mostly male names to be effectively related to males and mostly female names to be effectively related to females, and therefore we got 21,337 male values and 7,684 female values for a newly generated *Gender* attribute as a result of this first **data cleaning** process. As we will further discuss in Section 7.3, these decisions, together with the adoption of the `gender-guesser` library itself, represent some of the most critical choices we had to deal with, because they can lead to the introduction of technical bias, that is, as discussed in Section 2.1, bias originated from (or exacerbated by) design choices, constraints, and technological tools.

We also operated **data reduction** by removing the *Typical Hours*, *Hourly Rate*, and *First Name* columns, not relevant for our analysis.

As a consequence of the first data cleaning process, the number of different job titles decreased from 1,089 to 1,057. However, since the FAIR-DB tool used for bias analysis requires user interactions, and in order to lighten the

[3]Available at: `https://pypi.org/project/gender-guesser`.

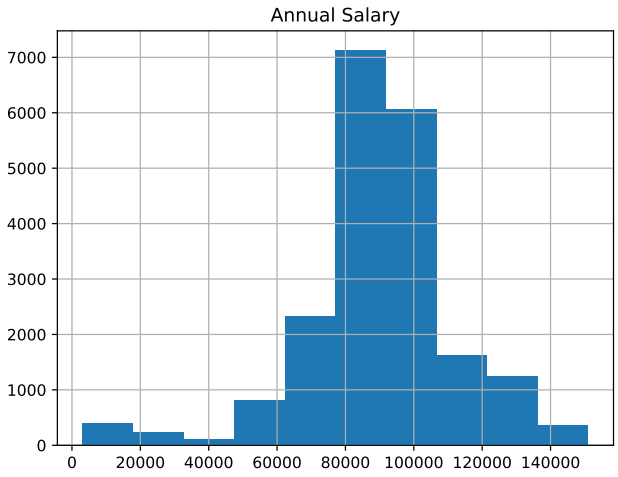


Figure 6.1: Distribution of the Annual Salary values for the Chicago dataset.

workload and speed up computational times, we decided to remove job titles with less than 100 occurrences.

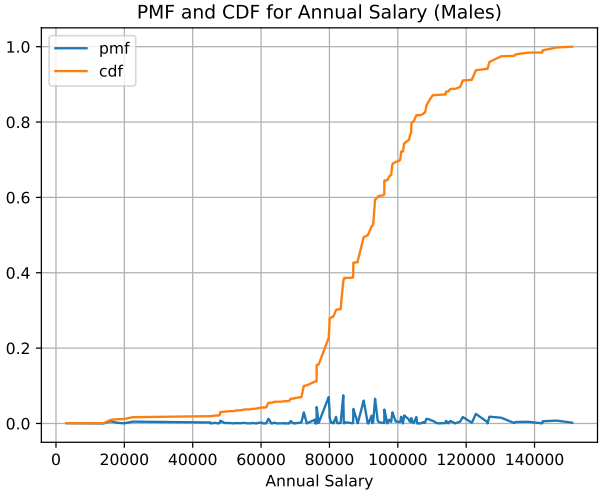
Our final preprocessed dataset includes 20,309 tuples, of which 16,146 males and 4,163 females, and with 35 distinct *Job Title* values and 20 distinct *Department* values.

We decided to plot the *Annual Salary* values distribution, in order to get a visual overview on the incomes and estimate possible threshold values for the creation of interval levels. The resulting graph is displayed in Figure 6.1.

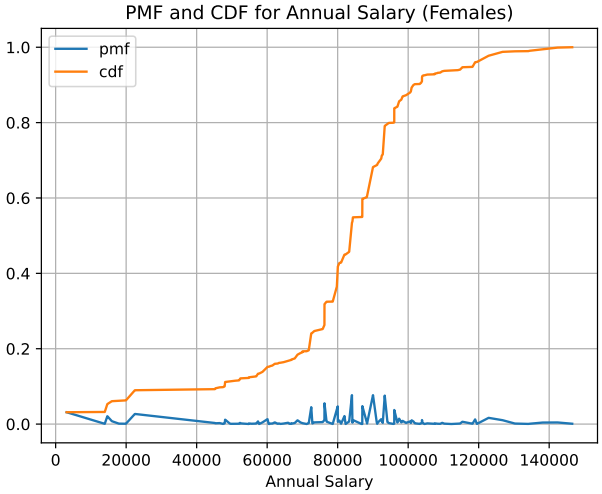
Lastly, we decided to plot the *Cumulative Distribution Function (CDF)* of male and female employees, for each *Annual Salary* value. For the laymen, CDF gives the probability of a discrete variable – *Annual Salary* in our case – to be less than or equal to a specific value. The comparison between Figure 6.2(a) and Figure 6.2(b) shows, once again, that women are more likely than men to earn less, and specifically to have an income in the range of [0, 40K] dollars per year.

6.1.3 The 'Glassdoor Method'

As already specified in Section 5.1, the point of reference for this method is the report published by Glassdoor in 2017 with the aim of helping HR practitioners in analyzing the internal gender pay gap of their companies [12]. Although the report provides a step-by-step guide for the statistical software R, we decided to use Python for our analysis, in order to better integrate the results with the ones from the other tools.



(a)



(b)

Figure 6.2: Cumulative distribution function of male (a) and female (b) employees, for each $Annual\ Salary$ *value of the Chicago dataset.*

The first step of the analysis, after cleaning up the data and loading them, consists in the creation of a couple of attributes useful for the statistical analysis: *Log Annual Salary* and *Male*. The former is simply the natural logarithm of the annual salary of the employee (i.e. the logarithm to the base of the mathematical constant e, approximately equal to 2.71828), useful because it provides a simple interpretation of the regression results; the latter

is a dummy indicator equal to 1 for males and 0 for females, which is the key variable of the analysis: if there is no gap, being male should not provide any advantage, and the coefficient of this variable in the regression should be equal to 0, otherwise its value would give us an estimate of the approximate percentage pay gap between men and women. It is worth to mention that the Glassdoor report also suggests to perform a discretization of age values of employees, grouping them into bins (25−, 25–34, 35–44, 45–54, 55+), but our dataset does not contain any information about the age of the employees.

The report suggests to look at the data before proceeding with the regressions, and it recommends to print a 'summary table' displaying the basic statistical information about the dataset. Figure 6.3(a) shows the table related to the Chicago dataset, and it displays the variables *Annual Salary*, *Log Annual Salary*, and *Male* sample size (count), arithmetic mean, minimum and maximum values, and standard deviation (i.e. a measure of the average amount of variability in the dataset – calculated as the square root of the variance – which tells, on average, how far each value lies from the mean). Another useful visualization tool is the so-called 'pivot table', displayed in Figure 6.3(b), which provides a high-level summary of the overall difference in pay between men and women by showing the arithmetic mean of the *Annual Salary* attribute values for males and females, together with the number of observations (len) and the median values (i.e. the numeric values separating the higher half of the samples from the lower half). The pivot table is also useful to get a first estimate of the 'unadjusted' pay gap: men on average are paid \$92,022.03 per year, while women on average earn \$79,790.83 per year – an overall 'unadjusted' pay gap of \$12,231.20 (13.3% of male pay). Lastly, since we are also interested in the 'adjusted' pay gap, it is important to look at the average salaries of men and women employed in the different job titles. Figure 6.3(c) shows the first 8 (out of 35) job titles in alphabetical order, displaying average salaries for men and women and sizes of the samples (i.e. number of men and women employed in the specific job title – information relevant to the problem of representation).

In order to estimate the gender pay gap, the reference linear regression model, as mentioned in Section 5.1, is:

$$LogAnnualSalary_i = \beta_1 Male_i + \beta_2 Controls_i + \epsilon_i$$

The report recommends to execute three different linear regressions: the first with no controls at all, regressing salary only on the male-female gender dummy (and therefore calculating the approximate overall percentage pay gap between men and women – the 'unadjusted' pay gap); the second with the addition of variables related to employee characteristics like highest

	Annual Salary	Log Annual Salary	Male
count	20309.00	20309.00	20309.00
mean	89514.84	11.35	0.80
std	22067.19	0.42	0.40
min	3120.00	8.05	0.00
max	151026.00	11.93	1.00

(a)

	average	median	len
	Annual Salary	Annual Salary	Annual Salary
Gender			
female	79790.83	84054.00	4163.00
male	92022.03	91338.00	16146.00

(b)

		average	len
		Annual Salary	Annual Salary
Job Title	Gender		
ADMINISTRATIVE ASST II	female	67908.82	102.00
	male	61292.00	9.00
AVIATION SECURITY OFFICER	female	73178.00	42.00
	male	73126.63	149.00
CAPTAIN-EMT	female	146538.00	4.00
	male	147328.84	160.00
CONSTRUCTION LABORER	female	92352.00	54.00
	male	92352.00	337.00
DETENTION AIDE	female	71730.48	50.00
	male	69382.60	131.00
ELECTRICAL MECHANIC	female	104000.00	12.00
	male	104000.00	194.00
FIRE ENGINEER-EMT	female	113901.86	14.00
	male	114411.09	344.00
FIREFIGHTER	female	102326.00	3.00
	male	102630.77	217.00

(c)

Figure 6.3: Summary table (a), pivot table (b) and average salaries of men and women employed in the different job titles (c) for the Chicago dataset.

education, years of experience, and performance evaluation scores; the third including all the possible controls (and finally estimating the 'adjusted' pay gap). Due to the lack of attributes, we performed only two linear regressions: the first with no controls and the second including *Job Title*, *Department*, and *Status*.

The results are shown in Figure 6.4: a coefficient of 0.242 on the male-female dummy variable means there is approximately 24.2% 'unadjusted'

Dependent Variable: Log Annual Salary		
	(1)	(2)
Male	0.242	0.004
Job Title		0.115
Department		0.726
Status		0.362
Constant	11.155	11.070
Controls		
- Job Title	No	Yes
- Department	No	Yes
- Status	No	Yes
Observations	20309	20309
R^2	0.053	0.950

Figure 6.4: Regression results for the Chicago dataset.

pay gap (therefore, men on average earn 24.2% more than women), but adding to the model all of the controls available in the data the coefficient value shrinks to 0.4% and becomes no longer statistically significant. In this case, we say there is no evidence of a systematic gender pay gap on an 'adjusted' basis, after controlling for observable differences between male and female workers, and the big discrepancy between the coefficient values is due to the overrepresentation of men in higher-paying roles and their underrepresentation in lower-paying jobs.

6.1.4 FAIR-DB

FAIR-DB is a tool based on functional dependencies, as already described in Section 5.2, and it operates by following the workflow shown in Figure 5.1.

- **Data preparation and exploration**: this phase is mostly covered by Section 6.1.2. In addition to the preprocessing techniques applied before, we had to deal with the **discretization** of *Annual Salary* values, since numbers are not really useful in estimating correlations between attributes (functional dependencies might otherwise involve really specific income values, and minimal differences among those values would not be perceived by the instrument, making it extremely difficult to detect occurrences and generate rules accordingly). We decided to create 2 interval levels (or bins) splitting *Annual Salary* values in $\leq$ 90K and $>$ 90K and generating a new *Annual Salary Bin* attribute to store this information, by following the approach presented

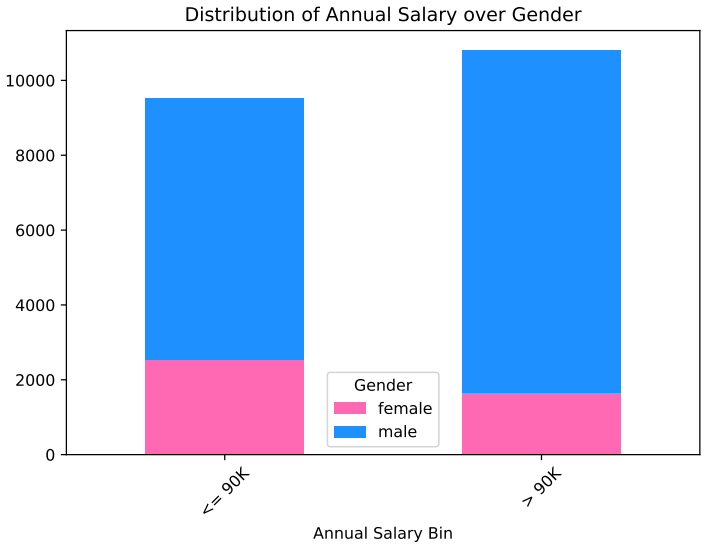


Figure 6.5: Distribution of the Annual Salary values for the Chicago dataset (2 bins).

in [5] by the authors of the tool. *Annual Salary Bin* represents our **target attribute**, while *Gender* is our **protected attribute**. The choice of 90K as threshold was made by looking at the *Annual Salary* values distribution, shown in Figure 6.1.

The histogram of Figure 6.5 shows instead the distribution of the annual salary of employees over their *Gender* attribute, and it is particularly useful because it provides a preliminary visual representation of the discrepancy between the number of male and female employees belonging to the same bin, highlighting the fact that, although the amount of people earning more than $90,000 is larger, there are many more women in the least profitable group.

Lastly, we performed a new **data reduction** operation by removing from the dataset the attributes *Name* and *Annual Salary*, not relevant anymore for our analysis, since the tool will make use of the *Annual Salary Bin* variable.

- **ACFD Discovery and filtering**: as specified in Section 5.2, this phase makes use of the *ACFD Discovery* algorithm, presented in [41]. The authors of the algorithm made available a compute capsule on Code Ocean[4], which works basically as a Web-based application, allowing the user to upload a CSV file (in our case, we exported and uploaded the

[4]Available at: `https://codeocean.com/capsule/6146641/tree`.

modified Chicago dataset), set the values of the required parameters (*minimum support*, *minimum confidence*, *maximum antecedent size*), run the algorithm and download the results in the form of a text file. The software also allows the choice of different algorithm implementations to be used to extract the dependencies but, as reported in the documentation, the default option FD-First-DFS-dfs is generally the fastest and there are no particular reasons, apart from performances, to choose one implementation over another one.

Since our dataset does not contain a large amount of attributes, we decided to keep *maximum antecedent size* = 2 (meaning that at most 2 variables will appear in the LHS of the computed rules). For what concerns the confidence, its value is computed as the ratio between the frequency of the dependency over the frequency of the LHS of the rule, and therefore a *minimum confidence* = 0.8 seemed to be a reasonable threshold, since the lower the parameter, the more dependencies are generated at each round increasing the computational complexity and making more difficult the subsequent choice of the most interesting ACFDs. Finally, we had to deal with the choice of a proper minimum support, which is quite a delicate operation: if it is too high the risk is to lose information about small groups, if it is too low there could be too many dependencies to analyze. We set *minimum support* = 100, because it is a reasonably low number if compared to the total number of tuples in the dataset and to the average amount of tuples for each *Job Title* value ($\frac{20{,}309}{35} \simeq 580$).

The text file generated by *ACFD Discovery*, containing 714 rules, has to be filtered, since the dependencies detected may not involve the protected attribute or the target attribute; there could also be dependencies in which some attribute values are not specified. The authors of [5] did not use the compute capsule of *ACFD Discovery*, and therefore were able to run the algorithm with an additional parameter, in order to discard the rules not containing the target attribute and its value. We balanced the gap by importing the text file, parsing it in order to extract every rule, and doing the same filtering operation a posteriori of *ACFD Discovery*. This operation resulted in a reduction in the number of dependencies from 714 to 145.

FAIR-DB then proceeds in filtering the rules by following 4 steps:

1. For each dependency, the LHS is separated from the RHS, and from both antecedent and consequent every couple ‘attribute -

```
Total number of tuples in df2:  49
                                                Rule  Supp  Conf  Diff  GenderDiff
0  {'lhs': {'Annual Salary Bin': '> 90K'},
    'rhs': {'Gender': 'male'}}                  0.45  0.85  0.05         NaN
1  {'lhs': {'Gender': 'female',
            'Salary or Hourly': 'Hourly'},
    'rhs': {'Annual Salary Bin': '<= 90K'}}     0.03  0.89  0.24        0.24
2  {'lhs': {'Annual Salary Bin': '> 90K',
            'Salary or Hourly': 'Hourly'},
    'rhs': {'Gender': 'male'}}                  0.07  0.94  0.13         NaN
3  {'lhs': {'Salary or Hourly': 'Salary',
            'Annual Salary Bin': '> 90K'},
    'rhs': {'Gender': 'male'}}                  0.38  0.83  0.04         NaN
4  {'lhs': {'Annual Salary Bin': '> 90K',
            'Gender': 'female'},
    'rhs': {'Salary or Hourly': 'Salary'}}      0.08  0.95  0.15        0.08
```

Figure 6.6: First 5 filtered dependencies with their metrics for the Chicago dataset (2 bins). NaN (Not a Number) means that the p-Difference has not been computed for the specific rule, since the protected attribute is not in the antecedent.

value' is stored.

2. For each rule, every couple 'attribute - value' is checked, and dependencies with missing values are discarded (being AFDs but not ACFDs).
3. A dictionary (that is, an unordered and indexed data structure similar to a list) of the remaining ACFDs is generated. It is made of two fields: 'lhs' and 'rhs', and each field contains a list of one or more couples 'attribute - value'.
4. Since each rule of the dictionary contains the target attribute but not necessarily any protected attribute, the dependencies are further parsed in order to satisfy both the criteria.

For the Chicago dataset, the filtering operation resulted in a reduction in the number of dependencies from 145 to 49. For each rule, the metrics *support*, *confidence*, *difference* and *p-Difference*, already introduced in Section 3.6 and Section 5.2, are computed, and the first occurrences are displayed in the form of a table, as shown in Figure 6.6.

- **ACFD selection**: in this phase FAIR-DB selects, among the filtered rules, the most 'unethical' according to the group fairness criterion, by looking at the computed metrics. It is worth to briefly recap the meanings behind the measures:

- **Support**: it expresses the percentage of records in the dataset that verifies the dependency – the higher the value, the more tuples are involved.
- **Confidence**: it shows how frequently the dependency is verified knowing that the antecedent is verified – the higher the value, the less approximate is the dependency.
- **Difference**: it indicates how much a dependency is 'unethical' – the higher the value, the more unfair is the dependency.
- **p-Difference**: it indicates how much the dependency shows bias paying attention to the specific value of a protected attribute – the higher the value, the more the rule is discriminatory with respect to the specific protected attribute value.

The selection of the most relevant rules takes place automatically, since the algorithm only keeps the dependencies with a difference parameter value higher than a minimum threshold imposed by the user. We decided to set *minimum difference* = 0.02, in order to keep the majority of the unfair dependencies, avoiding unintentionally ignoring some rules that could instead be relevant for us. This operation resulted in a reduction in the number of dependencies from 49 to 10.

After that, the algorithm performs what the authors call **ACFD completion**. Given the selected ACFDs, the framework computes all the possible combinations for each rule over the protected attributes and the target attribute (performing a Cartesian product between the attribute values). Taking as an example rule 0 of Figure 6.6:

$$AnnualSalaryBin = \text{`}> 90K\text{ '} \rightarrow Gender = \text{` male '}$$

we identify *Annual Salary Bin* as target attribute, whose possible values are $\leq$ 90K and $>$ 90K, and *Gender* as protected attribute, whose possible values are male and female. Therefore, the possible combinations for the rule are:

$$AnnualSalaryBin = \text{`}> 90K\text{ '} \rightarrow Gender = \text{` male '}$$

$$AnnualSalaryBin = \text{`}\leq 90K\text{ '} \rightarrow Gender = \text{` male '}$$

$$AnnualSalaryBin = \text{`}> 90K\text{ '} \rightarrow Gender = \text{` female '}$$

$$AnnualSalaryBin = \text{`}\leq 90K\text{ '} \rightarrow Gender = \text{` female '}$$

For each newly generated dependency, the evaluation metrics are computed, and a new automatic selection is performed, by keeping the

```
Number of original CFDs: 10
Number of combinations rules:  36
Number of final rules found:  18
                                          Rule  Supp  Conf  Diff  GenderDiff  Mean
0   {'lhs': {'Annual Salary Bin': '> 90K'},
     'rhs': {'Gender': 'male'}}               0.45  0.85  0.05         NaN  0.25
20  {'lhs': {'Status': 'F',
             'Annual Salary Bin': '> 90K'},
     'rhs': {'Gender': 'male'}}               0.45  0.85  0.04         NaN  0.25
16  {'lhs': {'Annual Salary Bin': '> 90K',
             'Gender': 'male'},
     'rhs': {'Salary or Hourly': 'Salary'}}   0.38  0.85  0.06       -0.01  0.22
12  {'lhs': {'Salary or Hourly': 'Salary',
             'Annual Salary Bin': '> 90K'},
     'rhs': {'Gender': 'male'}}               0.38  0.83  0.04         NaN  0.21
7   {'lhs': {'Gender': 'female',
             'Salary or Hourly': 'Hourly'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.03  0.89  0.24        0.24  0.14
```

Figure 6.7: First 5 selected and ranked dependencies with their metrics for the Chicago dataset (2 bins).

rules with difference greater than the respective minimum threshold. ACFD completion basically allows the user to study all the domain of the protected attributes and of the target class, and the computation of combinations brings to the surface also small groups that could not be studied otherwise. This operation, for the Chicago dataset, generated 36 dependencies (including the original 10), 18 of which have a difference above the threshold.

- **ACFD ranking**: the dependencies are ranked in descending order of support, difference, or mean, according to the user choice. As already mentioned in Section 5.2, the support option highlights the pervasiveness of the rules, the difference highlights their unethical aspect, and the mean represents the best trade-off between difference and support. Because of that, we decided to adopt the mean as ordering criterion. The resulting table is finally printed, and Figure 6.7 shows the first 5 dependencies (the others are displayed to the user but are omitted here for the sake of brevity).

- **ACFD user selection and scoring**: this last phase requires interaction from the user, who has to select N interesting dependencies among the ones previously ranked. The system then computes a final scoring outline based on 3 measures:

- **Cumulative support**: percentage of tuples of the dataset involved by the selected ACFDs – the higher the value, the more tuples are involved.
- **Difference mean**: arithmetic mean of all the 'Difference' columns of the selected ACFDs. It indicates how much the dataset is ethical according to the chosen rules – the higher the value, the more unfair is the dataset.
- **Protected attribute difference mean**: for each protected attribute, arithmetic mean of the p-Difference measure over all the selected ACFDs. It indicates how much the dataset is ethical over the protected attribute according to the chosen rules – the higher the value, the more the dataset is discriminatory with respect to the specific protected attribute.

For our research, we selected all the dependencies in which the target attribute appears in the RHS of the rule ($N = 6$ out of 18). This choice is due to the fact that the authors did not specify any criteria or suggestion for the manual selection of the rules, and the algorithm does not consider rules in which the protected attribute appears in the RHS as biased with respect to the protected attribute itself, in fact, as can be noticed in Figure 6.7, rules in which the protected attribute is in the RHS have a NaN (not computed) p-Difference value, while the same statistical measure is never NaN when the protected attribute is in the LHS. Other dependencies, in which both target attribute and protected attribute appear in the RHS, are not particularly relevant for us, precisely because none of our variables of interest is functionally dependent upon the other. We will further discuss about the impact of this choice in Section 7.3. The chosen ACFDs, together with the final scores, are displayed in Figure 6.8.

Among the chosen dependencies, we can detect a correspondence between rule 7 and rule 4: women paid on an hourly basis tend to earn less than \$90K, while men paid on an hourly basis tend to earn more than \$90K. These rules are the ones with the highest support (respectively 0.03 and 0.07), and rule 7 is the one with the highest difference value (0.24).

Even though the support is very low – and therefore not many tuples out of the total are involved – it is important to point out that rules 27 and 24 suggest a discriminatory behavior in the subgroup of the AVIATION department, while rules 29 and 30 show that, for what concerns the OEMC department, men seem to be less paid than women. The main reasons for

```
Number of tuples interested by the rules:  2310
Total number of tuples:  20309

Cumulative Support:  0.114
Difference Mean:  0.094
Gender Difference Mean:  0.094


Total number of ACFDs selected:  6
                                          Rule  Supp  Conf  Diff  GenderDiff  Mean
7   {'lhs': {'Gender': 'female',
             'Salary or Hourly': 'Hourly'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.03  0.89  0.24        0.24  0.14
27  {'lhs': {'Gender': 'female',
     'Department': 'AVIATION'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.01  0.92  0.13        0.13  0.07
4   {'lhs': {'Gender': 'male',
             'Salary or Hourly': 'Hourly'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.07  0.41  0.06        0.06  0.06
29  {'lhs': {'Gender': 'male',
             'Department': 'OEMC'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.01  0.94  0.08        0.08  0.05
24  {'lhs': {'Gender': 'male',
             'Department': 'AVIATION'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.01  0.23  0.03        0.03  0.02
30  {'lhs': {'Gender': 'female',
             'Department': 'OEMC'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.00  0.18  0.03        0.03  0.02
```

Figure 6.8: Final selected rules and scores for the Chicago dataset (2 bins).

these behaviors are a disproportion, in favor of male employees, in the number of men and women employed in the AVIATION department, and, on the contrary, a disproportion, in favor of female employees, in the number of men and women employed in the OEMC one. Disproportions can cause these situations when, within a dataset and for a specific *Job Title* value, there are some tuples with income above the threshold for both men and women but, for one of the two genders, the proportion of these tuples with respect to the total is very small compared to the other one. In this dataset, for example, for the OEMC department, the number of male employees earning more than \$90K is 9 (out of 147 men employed in OEMC), while the number of female employees earning more than \$90K is 73 (out of 414 women employed in OEMC). Although on average both men and women earn less than \$90K, the disproportion makes the algorithm perceive a discriminatory behavior towards male employees.

As for the scoring measures, a cumulative support of 0.114 means that

11.4% of the dataset is 'problematic' (2,310 tuples out of 20,309), while difference mean and gender difference mean (equal because in our dataset *Gender* is the only protected attribute) have a value of 0.094 because of the gap between the difference metric values of the selected ACFDs (above 0.1 for rule 27, above 0.2 for rule 7, below 0.1 for the other rules).

To conclude, we can say that the dataset seems to be quite fair with respect to the group fairness criterion, that is the one on which the tool is based, but more than 10% of the tuples show bias. Furthermore, the representation problem is not taken into account, and therefore the tendency of women to be employed in less profitable jobs than men, displayed in Figure 6.2 and Figure 6.6, is ignored.

6.1.5 Ranking Facts

As specified in Section 5.3, Ranking Facts is primarily meant to be a Web-based application with the aim of discovering fairness in a dataset by making use of ranking and providing to the user a collection of visual widgets. However, because of the size of our dataset, we could not use the tool in the form of a Web-based application, and we had to opt for the notebook version.

Before importing the dataset, we had to deal with a further **data transformation** process, in which we converted our categorical attributes (*Status*, *Job Title*, *Department*, *Salary or Hourly*) into numerical ones, since the tool can perform ranking only over them. A value of F for the *Status* attribute was therefore converted to 1, while P was converted to 0. The same holds for *Salary or Hourly* (Salary = 1, Hourly = 0), while for *Job Title* and *Department* numbers from 0 to 35 and from 0 to 20 respectively substituted the original categorical values.

Once imported the dataset, the tool initially plots the distribution of some attributes specified by the user (in our case, we decided to plot the distributions of the *Annual Salary* and *Gender* values). While the related distribution graphs are not particularly meaningful, the heatmap generated subsequently provides us with some preliminary information about the attribute correlations. As Figure 6.9 shows, there seems to be a significant correlation between *Job Title* and *Department*, but mostly important between *Status* and *Annual Salary*, *Status* and *Salary or Hourly*, and finally *Annual Salary* and *Salary or Hourly*, highlighting the fact that employees paid on an hourly basis and working part-time earn generally less than those paid on a salary basis and working full-time, and most of the part-time workers are also being paid on an hourly basis.

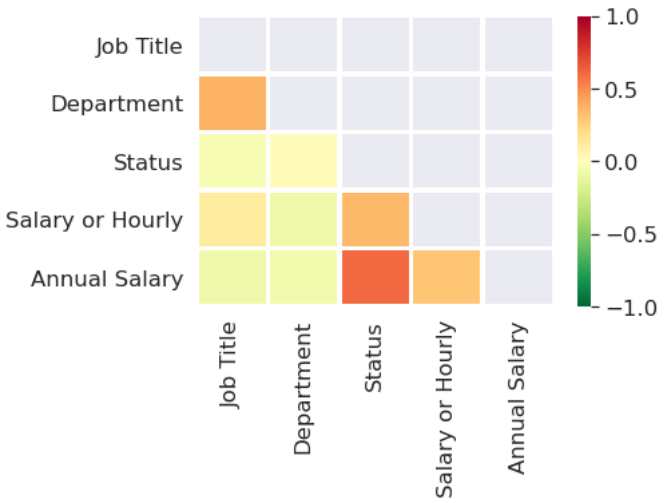


Figure 6.9: Heatmap showing attribute correlations for the Chicago dataset.

The tool then requires the user to specify some attributes to be used for the ranking, together with their weights. The reference formula is:

$$f(x) = w_1 \times Attribute_1(x) + \ldots + w_n \times Attribute_n(x)$$

In our case, the attributes used are *Job Title*, *Department*, *Status*, *Salary or Hourly*, and *Annual Salary*, because they are numerical (even though for *Job Title* and *Department* the numbering does not have an intrinsic meaning) and none of them is a protected variable. By following the examples provided by the authors of the tool, we decided to set the weights all equal to 1, giving the same importance to each attribute.

By following the widget description list provided in Section 5.3, we will now summarize our results.

- **Recipe** and **Ingredients**: the notebook version of the tool unfortunately does not provide any visual representation. As for the recipe, a couple of summary tables display some statistical measures (median, mean, minimum and maximum value) for the 5 attributes used in the ranking, respectively for the top-10 one and overall. These tables are not particularly useful, and neither is the information related to the ingredients, which tells us that the importance of each attribute used for the ranking is effectively equal to 1.

- **Stability**: this parameter explains whether the ranking methodology is robust on the specific dataset in use. Since an unstable ranking is one where slight changes to the data or to the methodology could lead

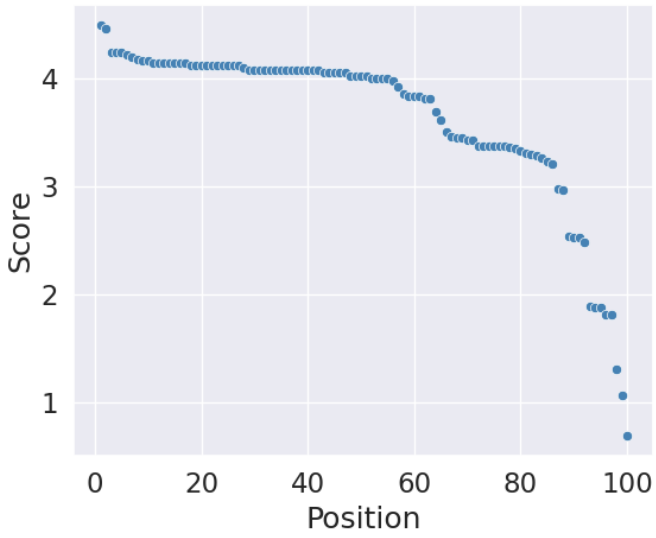
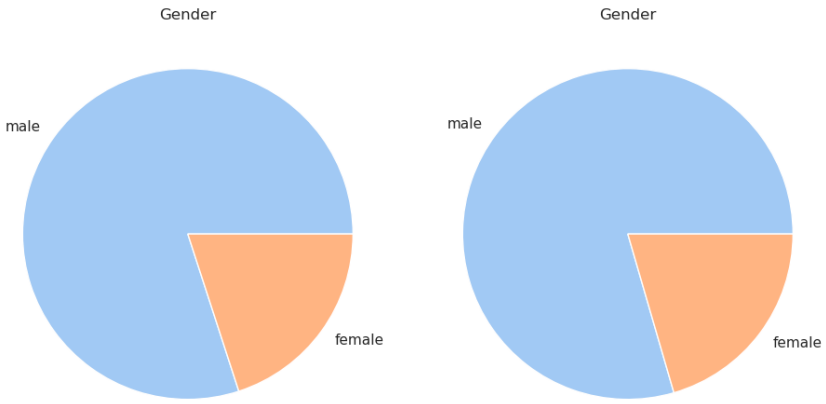


Figure 6.10: Score distribution of the ranking for the Chicago dataset.

to a significant change in the output, the label reports a stability score, as a single number that indicates the extent of the change required for the ranking to change. The stability of the ranking is quantified as the slope of the line that is fit to the score distribution, at the top-10 and overall. A score distribution is unstable if scores of items in adjacent ranks are close to each other ($|slope| \leq 0.25$), and so a very small change in scores will lead to a change in the ranking. The ranking used to analyze the Chicago dataset resulted to be unstable both at top-10 (stability at 0.08) and overall (stability at 0.0), and the score distribution is displayed in Figure 6.10.

- **Fairness**: it quantifies whether the ranked output exhibits statistical parity (group fairness) with respect to one or more protected attributes. The fairness measures adopted are all statistical tests in which the null hypothesis is that the ranking process is fair for the protected group, and whether a result is fair is determined by the computed p-value (a ranking is considered unfair when the p-value of the corresponding statistical test falls below 0.05).
  - **FA*IR**: the ranking resulted to be *fair for males*, with an approximate p-value of 0.99, and *fair for females*, with an approximate p-value of 0.32.
  - **Proportion**: the ranking resulted to be *fair for males*, with an approximate p-value of 1.0, and *fair for females*, with an



(a) *(b)*

Figure 6.11: Gender diversity widget for the top-10 (a) and overall (b) rankings of the Chicago dataset.

approximate p-value of 0.26.

– **Pairwise**: the ranking resulted to be *fair for males*, with an approximate p-value of 0.12, and *fair for females*, with an approximate p-value of 0.87.

The results seem therefore to be oriented towards a fair dataset, in accordance with the outcomes of the other tools.

- **Diversity**: it shows diversity with respect to a set of demographic categories of individuals, or a set of categorical attributes of other kinds of items, by displaying the proportion of each category in the top-10 ranked list and overall. Figure 6.11 shows the predominance of the male group over the female one, highlighting again a problem of gender representation.

6.2 Case Study 2: San Francisco

6.2.1 Dataset Description

The **San Francisco** dataset we considered includes 357,407 tuples and is made up of 10 attributes, briefly described as follows:

- *Employee Name*: full name of the employee in the form of 'Name Surname'.

- *Job Title*: categorical variable representing the job title of the employee (e.g. Firefighter). There are 2306 distinct values.
- *Base Pay*: numerical variable describing the annual regular pay for the employee.
- *Overtime Pay*: numerical variable describing the annual overtime pay for the employee.
- *Other Pay*: numerical variable describing other annual pay components for the employee.
- *Benefits*: numerical variable describing the amount of annual benefits for the employee.
- *Total Pay*: numerical variable describing the total annual salary of the employee, benefits excluded (*Base Pay* + *Overtime Pay* + *Other Pay*).
- *Total Pay + Benefits*: numerical variable describing the total annual salary of the employee, benefits included (*Base Pay* + *Overtime Pay* + *Other Pay* + *Benefits*).
- *Year*: numerical variable representing the year of reference (the dataset contains data related to the years 2011 to 2019).
- *Status*: binary categorical variable describing whether the employee is employed full-time (FT) or part-time (PT).

6.2.2 Data Preprocessing

As for the Chicago case, we operated some **data transformation** processes on the attributes of the San Francisco dataset. Specifically, the columns *Employee Name* and *Total Pay* were renamed respectively to *Name* and *Annual Salary*, and the attribute values for *Status* were transformed from FT and PT to simply F and P, in order to keep the algorithm used for the subsequent bias analysis as simple as possible and have a consistent structure across the datasets in use. We also operated a significant **data cleaning**, by filtering the tuples on the *Year* attribute, and keeping only the ones with *Year* = 2019, since they are the most recent and we want to avoid redundant data across multiple years. This operation resulted in a reduction in the number of tuples from 357,407 to 44,525.

Again, since the original dataset does not contain any gender-related information, we relied on `gender-guesser` to infer the gender of the employees from their *First Name*, by splitting the *Name* attribute and saving the

results in a newly generated *Gender* attribute. We obtained (out of the total of 44,525 tuples):

- unknown: 5,096 values.
- andy: 1,975 values.
- male: 20,636 values.
- female: 14,283 values.
- mostly_male: 1,153 values.
- mostly_female: 1,382 values.

We then removed, as we did for the Chicago dataset, tuples related to unknown and androgynous names, and we assumed mostly male names to be effectively related to males and mostly female names to be effectively related to females, and therefore we got 21,789 male values and 15,665 female values as a result of this **data cleaning** process.

We also operated **data reduction** by removing the *Base Pay*, *Overtime Pay*, *Other Pay*, *Benefits*, *Total Pay & Benefits*, *Year*, and *First Name* columns, since the information concerning the year became no longer useful and for the purpose of this research we are only interested in the total annual salary of employees.

Finally, we performed a last **data cleaning** process by removing job titles with less than 100 occurrences.

Our final preprocessed dataset includes 22,996 tuples, of which 13,688 males and 9,308 females, and with 81 distinct *Job Title* values.

Figure 6.12 shows the *Annual Salary* values distribution, from which we can observe that the lowest paid jobs are the most common, and the largest group of employees earns less than $50,000. This is due to the fact that, in comparison with the Chicago dataset, the San Francisco one contains many more part-time employees (9,308 out of 22,996 tuples for San Francisco, 639 out of 20,309 tuples for Chicago), and many job titles are monopolized or nearly monopolized by them, such as Pool Lifeguard, School Crossing Guard, or Recreation Leader.

Lastly, Figure 6.2(a) and Figure 6.2(b) display the Cumulative Distribution Function (CDF) of male and female employees, for each *Annual Salary* value, showing that the higher-paying jobs are done almost exclusively by men. Again, women are more likely than men to earn less, since the curve on the female graph increases more rapidly compared to the male one (representing a higher probability to have a lower income).

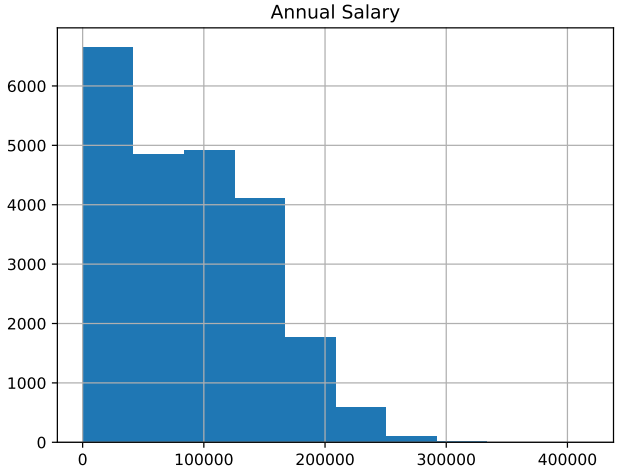


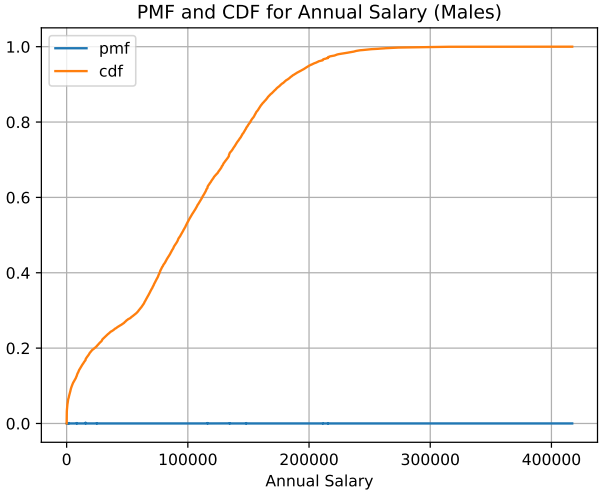
Figure 6.12: Distribution of the $Annual\ Salary$ *values for the San Francisco dataset.*

6.2.3 The 'Glassdoor Method'

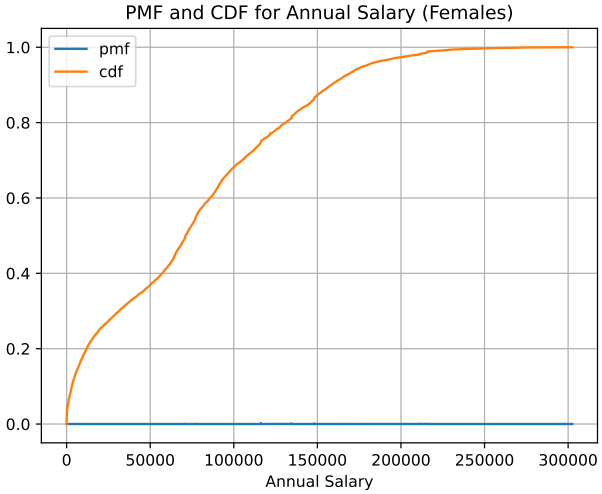
As for the Chicago case, we started by creating the *Log Annual Salary* and *Male* attributes, useful for the statistical analysis. After that, we proceeded in printing the summary and pivot tables, shown respectively in Figure 6.14(a) and Figure 6.14(b). Since men on average are paid \$94,370.73 per year, while women on average earn \$75,841.50 per year, we got a first estimate of the overall 'unadjusted' pay gap of \$18,529.23 (19.6% of male pay). Figure 6.14(c) shows the first 8 (out of 81) job titles in alphabetical order, displaying average salaries for men and women and sizes of the samples.

Again, because of the lack of attributes, we could perform only two linear regressions: the first with no controls and the second including *Job Title* and *Status*.

The results are shown in Figure 6.15: a coefficient of 0.304 on the male-female dummy variable means there is approximately 30.4% 'unadjusted' pay gap (therefore, men on average earn 30.4% more than women), but adding to the model all of the controls available in the data the coefficient value shrinks to -0.5% and becomes no longer statistically significant. We can conclude that also in this case there is no evidence of a systematic gender pay gap on an 'adjusted' basis, after controlling for observable differences between male and female workers, and again the big discrepancy between the coefficient values is due to the overrepresentation of men in higher-paying roles and their underrepresentation in lower-paying jobs.



(a)



(b)

Figure 6.13: Cumulative distribution function of male (a) and female (b) employees, for each $Annual\ Salary$ *value of the San Francisco dataset.*

6.2.4 FAIR-DB

- **Data preparation and exploration**: in addition to the preprocessing techniques applied in Section 6.2.2, we had once again to deal with the **discretization** of *Annual Salary* values and the creation of a new *Annual Salary Bin* attribute. By looking at the graph displayed in

	Annual Salary	Log Annual Salary	Male
count	22996.00	22996.00	22996.00
mean	86870.73	10.62	0.60
std	62399.93	2.02	0.49
min	0.01	-4.61	0.00
max	417152.63	12.94	1.00

(a)

	average	median	len
	Annual Salary	Annual Salary	Annual Salary
Gender			
female	75841.50	71100.98	9308.00
male	94370.73	94133.26	13688.00

(b)

		average	len
		Annual Salary	Annual Salary
Job Title	Gender		
Administrative Analyst	female	83282.16	89.00
	male	81709.54	102.00
Assistant Engineer	female	99238.61	63.00
	male	103185.83	100.00
Assoc Engineer	female	123409.06	46.00
	male	125359.07	123.00
Attorney (Civil/Criminal)	female	154756.46	190.00
	male	162993.18	199.00
Automotive Mechanic	female	96526.73	1.00
	male	98182.72	159.00
Automotive Service Worker	female	94050.96	3.00
	male	77288.95	123.00
Behavioral Health Clinician	female	77599.23	123.00
	male	81235.70	33.00
Camp Assistant	female	3897.82	60.00
	male	3949.49	48.00

(c)

Figure 6.14: Summary table (a), pivot table (b) and average salaries of men and women employed in the different job titles (c) for the San Francisco dataset.

Figure 6.12, we decided to keep 90K as threshold value, and therefore to use the same bins ($\leq$ 90K and $>$ 90K) used for the Chicago case.

The histogram of Figure 6.16 shows the distribution of the annual salary of employees over their *Gender* attribute. As we can notice, although the number of male and female employees belonging to the $\leq$ 90K bin is comparable, almost $\frac{2}{3}$ of the employees belonging to the $>$ 90K bin are men.

Dependent Variable: Log Annual Salary		
	(1)	(2)
Male	0.304	-0.050
Job Title		0.108
Status		0.349
Constant	10.443	11.618
Controls		
- Job Title	No	Yes
- Status	No	Yes
Observations	22996	22996
R^2	0.005	0.476

Figure 6.15: Regression results for the San Francisco dataset.

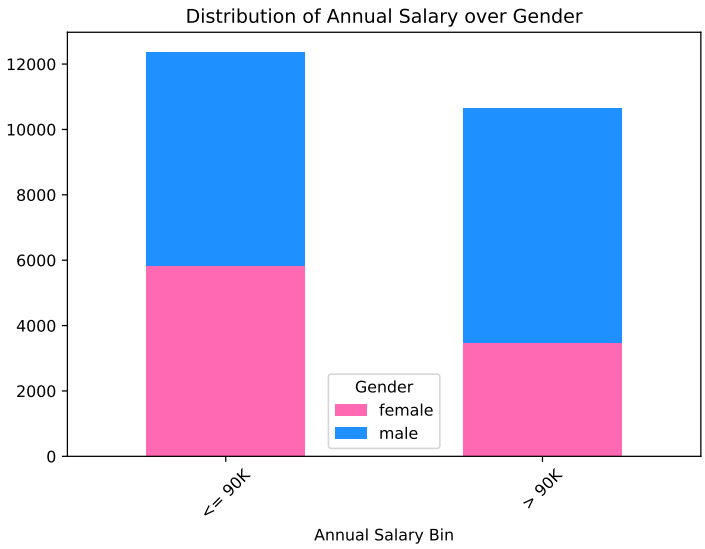


Figure 6.16: Distribution of the Annual Salary values for the San Francisco dataset (2 bins).

Lastly, we performed a new **data reduction** operation by removing from the dataset the attributes *Name* and *Annual Salary*, since the tool will make use of the *Annual Salary Bin* variable.

- **ACFD Discovery and filtering**: we used the compute capsule of the *ACFD Discovery* algorithm as we did for the Chicago case, and we decided to keep the same parameter values: *maximum antecedent size* = 2 because the dataset does not contain many attributes, *minimum confidence* = 0.8 to generate not too high a number of dependencies, and

minimum support = 100 because it is a reasonably low number if compared to the total number of tuples in the dataset and to the average amount of tuples for each *Job Title* value ($\frac{22{,}996}{81} \simeq 284$).

The text file generated by *ACFD Discovery*, containing 586 dependencies, was then imported, parsed, and filtered in order to discard the rules not containing the target attribute and its value. This operation resulted in a reduction in the number of dependencies from 586 to 220. The automatic filtering operations performed by FAIR-DB further reduced this number from 220 to 61.

- **ACFD selection**: as for the previous phase, we decided to maintain the same value for the only parameter required by ACFD selection, and therefore we set *minimum difference* = 0.02 in order to keep the majority of the unfair dependencies. The first automatic selection resulted in a reduction in the number of dependencies from 61 to 10. The subsequent ACFD completion and further selection generated 36 dependencies (including the original 10), 16 of which have a difference above the threshold.

- **ACFD ranking**: Figure 6.17 shows the first 5 dependencies, ranked according to the mean (the others are displayed to the user but are omitted here for the sake of brevity).

- **ACFD user selection and scoring**: for the reasons already specified in Section 6.1.4, we selected all the dependencies in which the target attribute appears in the RHS of the rule ($N = 10$ out of 16). The chosen ACFDs, together with the final scores, are displayed in Figure 6.18.

Among the chosen dependencies, we can detect a correspondence between rule 1 and rule 2: apparently, female part-time workers tend to earn more than \$90K, while male part-time workers tend to earn less than \$90K. These rules are the ones with the highest support (respectively 0.19 and 0.02). By looking at the average salaries of men and women employed in the different job titles, we found a few reasons for this behavior: first of all, in general there are many more part-time female employees out of the total number of women in the dataset, in comparison to men (4,162 out of 9,308 tuples for women, 4,606 out of 13,688 tuples for men); secondly, there are some typically masculine jobs (e.g. Automotive Mechanic, Electronic Maintenance Tech, Stationary Engineer) in which just a few women are employed, as full-time workers, while men are much more numerous and spread among full-time and part-time positions, and for these jobs the part-time income is on average

```
Number of original CFDs: 10
Number of combinations rules:  36
Number of final rules found:  16
                                        Rule  Supp  Conf  Diff  GenderDiff  Mean
4   {'lhs': {'Annual Salary Bin': '> 90K',
             'Gender': 'male'},
     'rhs': {'Status': 'F'}}                  0.30  0.96  0.36        0.04  0.33
5   {'lhs': {'Annual Salary Bin': '> 90K',
             'Gender': 'female'},
     'rhs': {'Status': 'F'}}                  0.13  0.84  0.24       -0.08  0.18
1   {'lhs': {'Gender': 'male',
             'Status': 'P'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.19  0.93  0.03        0.03  0.11
32  {'lhs': {'Annual Salary Bin': '> 90K',
             'Job Title': 'Transit Operator'},
     'rhs': {'Gender': 'male'}}               0.03  0.86  0.06         NaN  0.05
19  {'lhs': {'Gender': 'female',
             'Job Title': 'Assoc Engineer'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  0.17  0.09        0.09  0.05
```

Figure 6.17: First 5 selected and ranked dependencies with their metrics for the San Francisco dataset (2 bins).

lower than the threshold. On the contrary, there are some typically feminine jobs (e.g. Medical Evaluations Assistant, Nurse Practitioner, Registered Nurse) in which just a few men are employed, as full-time workers, while women are much more numerous and spread among full-time and part-time positions, and for these jobs the part-time income is on average higher than the threshold.

Even though the support is very low – and therefore not many tuples out of the total are involved – it is important to point out that rules 19 and 16 suggest a discriminatory behavior in the subgroup of the Assoc Engineer job title in favor of men (with the highest difference value for rule 19), and the same holds for rules 23 and 20 in the subgroup of assistant engineers. The main reason for these behaviors is a disproportion, in favor of male employees, in the number of men and women employed in these profitable professions.

Although the complementary rules have not been selected because of a low difference value, rule 12 shows that male registered nurses earn more than the threshold, and rule 11 shows that, for what concerns parking control officers, women tend to have an income lower than \$90K. Finally, rules 25 and 26 show that, for the HSA Sr Eligibility Worker job title, men seem to be less paid than women.

As for the scoring measures, a cumulative support of 0.243 means that

```
Number of tuples interested by the rules:  5596
Total number of tuples:  22996

Cumulative Support:  0.243
Difference Mean:  0.037
Gender - Difference Mean:  0.037


Total number of ACFDs selected:  10
                                        Rule  Supp  Conf  Diff  GenderDiff  Mean
1   {'lhs': {'Gender': 'male',
             'Status': 'P'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.19  0.93  0.03        0.03  0.11
19  {'lhs': {'Gender': 'female',
             'Job Title': 'Assoc Engineer'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  0.17  0.09        0.09  0.05
25  {'lhs': {'Gender': 'male',
             'Job Title': 'HSA Sr Eligibility Worker'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  0.92  0.06        0.06  0.03
2   {'lhs': {'Gender': 'female',
             'Status': 'P'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.02  0.12  0.03        0.03  0.03
16  {'lhs': {'Gender': 'male',
             'Job Title': 'Assoc Engineer'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.01  0.95  0.03        0.03  0.02
23  {'lhs': {'Gender': 'female',
             'Job Title': 'Assistant Engineer'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  0.16  0.04        0.04  0.02
12  {'lhs': {'Gender': 'male',
             'Job Title': 'Registered Nurse'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.01  0.89  0.02        0.02  0.02
11  {'lhs': {'Gender': 'female',
             'Job Title': 'Parking Control Officer'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  0.82  0.02        0.02  0.01
20  {'lhs': {'Gender': 'male',
             'Job Title': 'Assistant Engineer'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.00  0.90  0.02        0.02  0.01
26  {'lhs': {'Gender': 'female',
             'Job Title': 'HSA Sr Eligibility Worker'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.00  0.16  0.02        0.02  0.01
```

Figure 6.18: Final selected rules and scores for the San Francisco dataset (2 bins).

24.3% of the dataset is 'problematic' (5,596 tuples out of 22,996), while difference mean and gender difference mean have a value of 0.037 because each rule has a quite low value for the difference metric (below 0.1).

To conclude, we can say that the dataset seems to be quite fair with respect to the group fairness criterion, because even if almost 25% of the tuples seem to be biased, each rule has a low value both for the difference

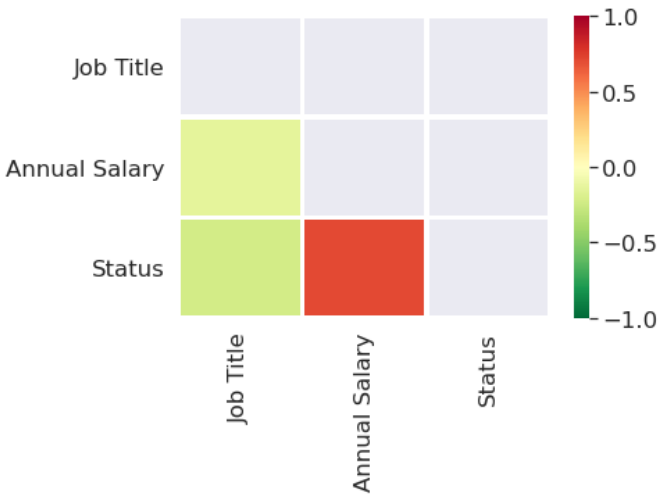


Figure 6.19: Heatmap showing attribute correlations for the San Francisco dataset.

metric – which, as already specified, determines the 'unfairness level' – and for the support one – indicating a low percentage of tuples involved. The most impacting dependency is indeed the one regarding male part-time employees, with a support value of 0.19. Furthermore, it is interesting to notice that for traditionally higher-paying jobs – in this case two branches of engineering – men seem to have an economic advantage over women, while the opposite condition occurs only in situations in which the female presence is typically much more numerous than the male one – that is, in our scenario, the HSA Sr Eligibility Worker job title and the part-time worker status.

6.2.5 Ranking Facts

As for the Chicago case, because of the size of our dataset, we could not use Ranking Facts in the form of a Web-based application, and we had to opt for the notebook version. Before importing the dataset, we operated a **data transformation** process, in which the categorical attributes *Job Title* and *Status* were converted into numerical ones.

Figure 6.19 shows the heatmap related to our dataset, and it suggests a really strong correlation between the attributes *Status* and *Annual Salary*, highlighting the fact that part-time employees tend to earn less than full-time employees.

Because of the lack of attributes, we could only use *Job Title*, *Status*, and *Annual Salary* as ranking parameters, all with a weight of 1 as done in the Chicago case and by following the examples provided by the authors.

The following list summarizes our results by following the widget descrip-

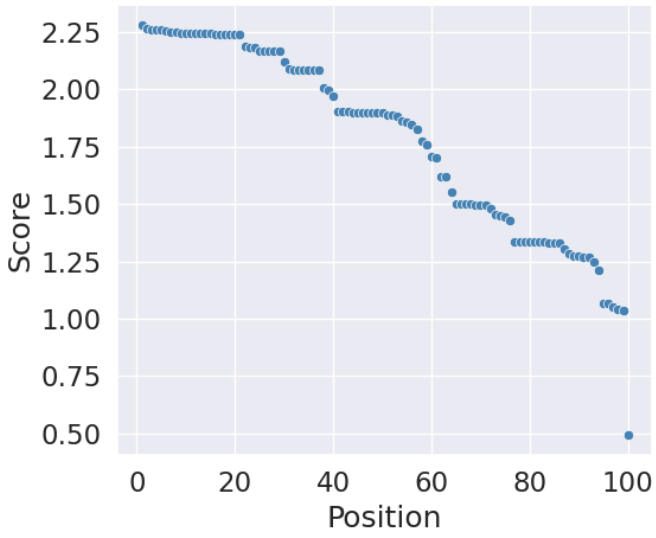
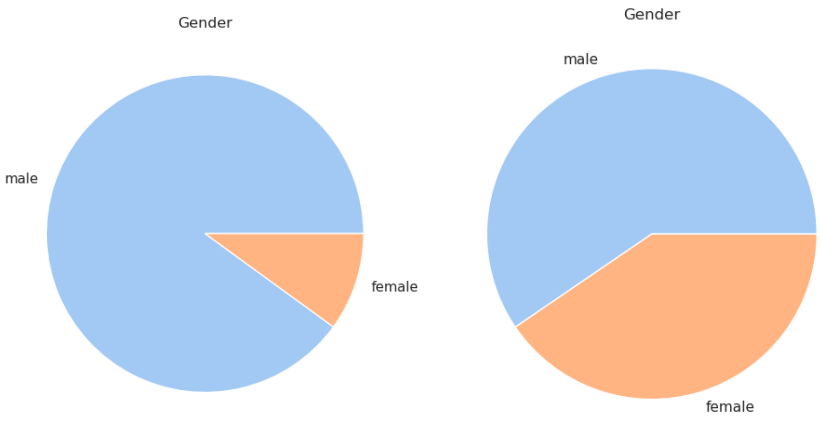


Figure 6.20: Score distribution of the ranking for the San Francisco dataset.

tion list provided in Section 5.3:

- **Recipe** and **Ingredients**: as for the Chicago case, these widgets did not provide us any particularly useful information. For the sake of completeness, we just report that the importance of each attribute used for the ranking is effectively equal to 1.
- **Stability**: our ranking resulted to be unstable both at top-10 (stability at 0.07) and overall (stability at 0.0). The score distribution is displayed in Figure 6.20.
- **Fairness**: recalling the fact that, for the statistical measures adopted, a ranking is considered unfair when the p-value of the corresponding test falls below 0.05, we will now recapitulate our fairness results:
  - **FA*IR**: the ranking resulted to be *fair for males*, with an approximate p-value of 1.0, and *unfair for females*, with an approximate p-value of 0.0.
  - **Proportion**: the ranking resulted to be *fair for males*, with an approximate p-value of 1.0, and *unfair for females*, with an approximate p-value of 0.0.
  - **Pairwise**: the ranking resulted to be *fair for males*, with an approximate p-value of 1.0, and *unfair for females*, with an approximate p-value of 0.0.



(a) (b)

Figure 6.21: Gender diversity widget for the top-10 (a) and overall (b) rankings of the San Francisco dataset.

The results seem to be oriented towards a biased dataset in favor of men, in contrast with the generally fair outcomes of the other tools. Following an analysis of the code and some tests, we then concluded that the proportion of women in the top-k results to be effectively very low, even bringing k from the standard value used in the algorithm (100) to 1000.

- **Diversity**: Figure 6.21 shows the predominance of the male group over the female one, especially in the top-10 ranking, highlighting again a problem of gender representation.

6.3 Other Design Choices

The aim of this section is to document in a technical way some experiments made on the datasets previously analyzed – Chicago and San Francisco – in which different design decisions were taken. It is worth to emphasize once more the importance of human choices behind computer systems, in particular when dealing with sensitive concepts such as fairness, and we believe that by looking at the impact of other design decisions we can get a broader perspective on how the tools should be used. Because of the technical nature of this chapter, the socio-ethical impact of these choices, as well as the impact of other ones not mentioned in this section, will not be discussed here but in Chapter 7.

6.3.1 Part-Time Employees Removal

One of the first design choices we had to deal with was concerning part-time employees. The Glassdoor report [12] indeed recommends to only include full-time employees in the gender pay gap analysis (or in case of equal amounts of full-time and part-time employees to conduct two separate analyses) because of the big differences between full-time and part-time workers in the labor market. We initially decided to exclude part-time employees from the datasets, by removing their tuples and subsequently also the *Status* attribute. This choice resulted to be penalizing in both our cases for different reasons:

- **Chicago**: for the preprocessed dataset, the number of male employees decreased from 16,146 to 15,880 (−1.65%), while the number of female employees decreased from 4,163 to 3,790 (−8.96%). Even though the reduction in the number of tuples is not excessive, most of the records removed are related to women, already underrepresented when compared to the number of men.

- **San Francisco**: for the preprocessed dataset, the number of male employees decreased from 13,688 to 9,082 (−33.65%), while the number of female employees decreased from 9,308 to 4,696 (−49.55%). Even if the number of male and female employees removed is similar, we found the removal of 9,308 tuples from the dataset to have too much impact on the dataset itself.

Furthermore, it is important to note that the Glassdoor report is meant to be a guide for HR practitioners in analyzing the internal gender gap of a company, while we are performing our analysis not on a company but on public employees of different sectors. Lastly, the removal of the *Status* variable further penalizes datasets already characterized by a low number of attributes, making it more difficult to get concrete results from the tools. Because of these reasons, we decided to include again part-time employees in the datasets and to proceed in our analysis as discussed before.

6.3.2 FAIR-DB: Discretization Using More Bins

For both the Chicago and the San Francisco datasets, we decided to conduct a FAIR-DB analysis using more than just 2 bins. Therefore, we split the *Annual Salary* values in 8 different interval levels (the K specification is implied): 0–39, 40–59, 60–79, 80–99, 100–119, 120–139, 140+.

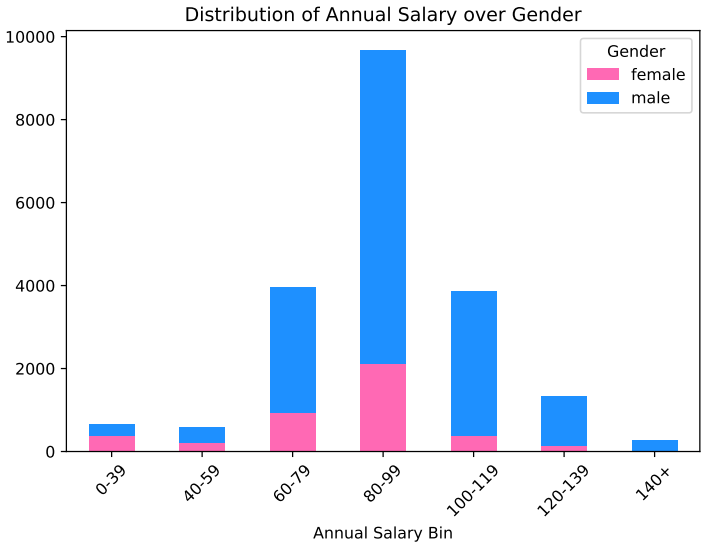


Figure 6.22: Distribution of the Annual Salary values for the Chicago dataset (8 bins).

We will now briefly summarize our results, without going too much into the details of the algorithm but describing the most relevant differences in comparison with the 2-bin analysis.

- **Chicago**: the preprocessed dataset is the same used for the 2-bin analysis, with 20,309 tuples of which 16,146 males and 4,163 females. The histogram of Figure 6.22 shows the distribution of the annual salary of employees over their *Gender* attribute. As we can notice, most women are concentrated in the 80–99 bin, which is the most numerous overall, but the proportion between men and women in lower bins is unbalanced: more than 50% of the employees in the 0–39 bin are females, even though women represent about $\frac{1}{5}$ of the population of the dataset.

  Since the dataset is the same as that used for the 2-bin analysis, we ran the *ACFD Discovery* algorithm keeping the same parameter values: *maximum antecedent size* = 2, *minimum confidence* = 0.8, and *minimum support* = 100. The algorithm generated 759 dependencies, and the number shrank to 193 when we filtered the results keeping only the rules containing the target attribute and its value. The automatic filtering operations performed by FAIR-DB further reduced this number from 193 to 64.

  We decided to keep *minimum difference* = 0.02 during the ACFD selection phase, in order to be able to compare the results of the two

```
Number of tuples interested by the rules:  86
Total number of tuples:  20309

Cumulative Support:  0.004
Difference Mean:  0.124
Gender - Difference Mean:  0.124


Total number of ACFDs selected:  2
                                          Rule  Supp  Conf  Diff  GenderDiff  Mean
156 {'lhs': {'Gender': 'male',
             'Job Title': 'ADMINISTRATIVE ASST II'},
      'rhs': {'Annual Salary Bin': '40-59'}}  0.00  0.44  0.23        0.23  0.11
161 {'lhs': {'Gender': 'female',
             'Job Title': 'ADMINISTRATIVE ASST II'},
      'rhs': {'Annual Salary Bin': '60-79'}}  0.00  0.80  0.02        0.02  0.01
```

Figure 6.23: Final selected rules and scores for the Chicago dataset (8 bins).

analyses and evaluate the impact of the different discretization processes performed. The first automatic selection resulted in a reduction in the number of dependencies from 64 to 28. The subsequent ACFD completion and further selection generated 172 dependencies (including the original 28), 66 of which have a difference above the threshold.

Lastly, we applied the same criterion used in the previous cases for manually choosing the rules (target attribute in the RHS). Unfortunately, we only got $N = 2$ out of 66 dependencies. The chosen ACFDs, together with the final scores, are displayed in Figure 6.23.

As we can see, both rules refer to the ADMINISTRATIVE ASST II job title, and they suggest a discriminatory behavior in favor of women, who seem to be more paid for this specific role (even though the support of 0.0 indicates a really low percentage of tuples involved). This situation did not show up in the 2-bin analysis, because even if apparently there is a gap between salaries of men and women, all the incomes fall below the threshold of \$90K, and therefore no dependency could have been generated. On the other side, the discriminatory behaviors detected in the 2-bin analysis, related to the AVIATION and OEMC departments, and to female employees paid on an hourly basis, were not detected here, presumably because the annual salary values reside all in the 80–99 bin.

- **San Francisco**: the preprocessed dataset is the same used for the 2-bin analysis, with 22,996 tuples of which 13,688 males and 9,308

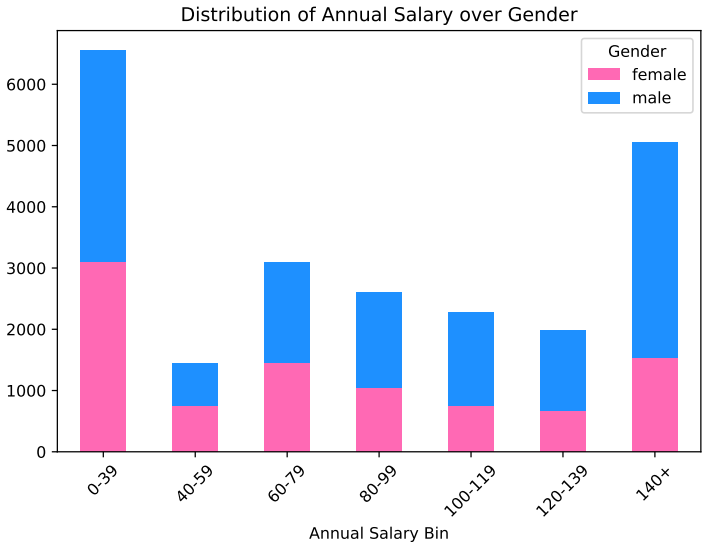


Figure 6.24: Distribution of the Annual Salary values for the San Francisco dataset (8 bins).

females. The histogram of Figure 6.24 shows the distribution of the annual salary of employees over their *Gender* attribute. We can see that the ratio of the number of men to women tends to increase: in lower bins there is a balanced proportion of male and female employees, while for higher-paying jobs men are much more numerous than women.

Again, we kept *maximum antecedent size* = 2, *minimum confidence* = 0.8, and *minimum support* = 100 for *ACFD Discovery*. The algorithm generated 465 dependencies, and the number shrank to 144 when we filtered the results keeping only the rules containing the target attribute and its value. The automatic filtering operations performed by FAIR-DB further reduced this number from 144 to 40.

By keeping *minimum difference* = 0.02, the first automatic ACFD selection resulted in a reduction in the number of dependencies from 40 to 20. The subsequent ACFD completion and further selection generated 154 dependencies (including the original 20), 62 of which with a difference above the threshold.

At the end, we got $N = 6$ out of 62 dependencies in the ACFD user selection phase. The chosen ACFDs, together with the final scores, are displayed in Figure 6.25.

The 6 final rules obtained are quite peculiar because they refer just to 2 different job titles. For what concerns junior clerks, men seem to

```
Number of tuples interested by the rules:  227
Total number of tuples:  22996

Cumulative Support:  0.010
Difference Mean:  0.037
Gender - Difference Mean:  0.037


Total number of ACFDs selected:  6
                                          Rule  Supp  Conf  Diff  GenderDiff  Mean
51  {'lhs': {'Gender': 'female',
             'Job Title': 'Engineer'},
     'rhs': {'Annual Salary Bin': '120-139'}}  0.00  0.14  0.05        0.05  0.03
40  {'lhs': {'Gender': 'female',
             'Job Title': 'Junior Clerk'},
     'rhs': {'Annual Salary Bin': '40-59'}}    0.00  0.15  0.04        0.04  0.02
32  {'lhs': {'Gender': 'male',
             'Job Title': 'Junior Clerk'},
     'rhs': {'Annual Salary Bin': '0-39'}}     0.00  0.92  0.04        0.04  0.02
43  {'lhs': {'Gender': 'male',
             'Job Title': 'Engineer'},
     'rhs': {'Annual Salary Bin': '140+'}}     0.00  0.89  0.03        0.03  0.02
53  {'lhs': {'Gender': 'female',
             'Job Title': 'Engineer'},
     'rhs': {'Annual Salary Bin': '0-39'}}     0.00  0.06  0.04        0.04  0.02
54  {'lhs': {'Gender': 'female',
             'Job Title': 'Engineer'},
     'rhs': {'Annual Salary Bin': '40-59'}}    0.00  0.03  0.02        0.02  0.01
```

Figure 6.25: Final selected rules and scores for the San Francisco dataset (8 bins).

be less paid then women. For the Engineer job title instead, women tend to earn less, and their incomes vary widely, since they fall within 3 different interval levels (0–39, 40–59, 120–139), while male engineers are paid more than \$140K. As for the Chicago case, the 2-bin analysis could not detect any of these discriminatory behaviors, because in the former case the salary values are all lower than the threshold, while in the latter case the presence of female engineers earning between \$120K and \$139K 'balanced' the gap, eluding the algorithm. For the same reasons, none of the discriminatory behaviors detected through the 2-bin analysis has been observed here.

6.3.3 FAIR-DB: Choice of Different Dependencies

As already specified, the ACFD user selection phase of the FAIR-DB framework is a delicate step, because the user needs to manually select N among all the dependencies for the subsequent scoring and calculation of the metrics.

In Section 6.1.4 we clarified the reasons why we decided to always keep rules in which the target attribute is in the RHS. However, during our first experiments, we tried to select all the dependencies, in order to check the impact of the user choices on the final results. The experiments were conducted on the same preprocessed datasets, and using the same parameter values; the only difference is indeed the selection of every rule instead of just some of them. The following results refer to the 8-bin analysis:

- **Chicago**: by selecting all 66 dependencies, we got a cumulative support of 0.856, meaning that 85.6% of the dataset is 'problematic' (17,384 tuples out of 20,309). The difference mean value is 0.153, while the gender difference mean is equal to 0.071. The measures have different values because for most rules (of which all with the target attribute in the LHS) the p-Difference is NaN, while the difference is not.
- **San Francisco**: by selecting all 62 dependencies, we got a cumulative support of 0.925, meaning that 92.5% of the dataset is 'problematic' (21,279 tuples out of 22,996). The difference mean value is 0.153, while the gender difference mean is equal to 0.003.

We can conclude that the user choices in this phase strongly impact the final outcomes, since a dataset results to be fair or not depending on the selected rules. In both our cases, the gap is huge: for Chicago the percentage of 'problematic' tuples shifted from 0.4% (as we can see from the cumulative support in Figure 6.23) to 85.6%, while for San Francisco the same percentage shifted from 1% (Figure 6.25) to 92.5%.

6.3.4 Grouping of Job Titles

Another recommendation from the Glassdoor report [12] is the one of grouping together similar job titles, because having too many unique roles with just a few workers in each could make the analysis less reliable. Therefore, we decided to test our tools on a slightly different version of the Chicago dataset, in which the data preprocessing phase consists of one more step: a **generalization** of the *Job Title* values. We opted for the Chicago dataset instead of the San Francisco one mostly because a certain degree of precision is required in classifying job titles, and since the classification has to be performed manually we preferred to deal with 35 distinct *Job Title* values rather than 81. In order to group job titles in a sensible way, we relied on a document published by the Equality Commission for Northern Ireland in 2013 [23]. Even though the document refers to Northern Ireland, it provides an exhaustive list of thousands of different job titles, together

with a few introductory sections on how to use the index, and we found useful to rely on these guidelines in grouping the job titles of our dataset. The document, based on the Standard Occupational Classification 2010 (SOC2010) – a common classification of occupational information for the U.K. – distinguishes between 9 different major groups:

1. Managers and senior officials
2. Professional occupations
3. Associate professional and technical occupations
4. Administrative and secretarial occupations
5. Skilled trades occupations
6. Personal service occupations
7. Sales and customer service occupations
8. Process, plant and machine operatives
9. Elementary occupations

The index then provides correspondences between each job title and the major group to which it belongs. Figure 6.26 shows the correspondences between our job titles and the related major groups, together with the index entries of reference. It is worth to mention that we decided to overwrite the previous *Job Title* values with the corresponding major group numbers, transforming the categorical attribute into numeric.

We will now provide the results of the analysis of the modified dataset:

- **The 'Glassdoor Method'**: we detected a bigger gap (in favor of men) in the average salaries of male and female employees for a few job title values. In particular, the most significant differences are related to job title 2 (professional occupations), in which the average income for men is \$102,539.79 while for women is \$73,827.52, and job title 6 (personal service occupations), in which the average income for men is \$75,868.14 while for women is \$42,134.53.

  These differences slightly impacted the *Male* coefficient of the 'adjusted' pay gap, bringing its value from 0.004 (standard Chicago dataset) to 0.034, meaning that men on average earn 3,4% more than women. The result, however, remains not statistically significant, and we cannot infer the presence of a systematic gender pay gap.

```
'ADMINISTRATIVE ASST II':            4 (Assistant, administrative)
'AVIATION SECURITY OFFICER':         9 (Officer, security)
'CAPTAIN-EMT':                       1 (Captain, fire)
'CONSTRUCTION LABORER':              9 (Worker, construction)
'DETENTION AIDE':                    6 (Aide, ward)
'ELECTRICAL MECHANIC':               5 (Mechanic, electrical)
'FIRE ENGINEER-EMT':                 3 (Engineer, fire)
'FIREFIGHTER':                       3 (Firefighter)
'FIREFIGHTER-EMT':                   3 (Firefighter)
'FIREFIGHTER-EMT (RECRUIT)':         3 (Firefighter)
'FIREFIGHTER/PARAMEDIC':             3 (Paramedic-ECP)
'FOSTER GRANDPARENT':                6 (Parent, foster)
'GENERAL LABORER - DSS':             9 (Operative, cleansing (street cleaning))
'HOISTING ENGINEER':                 5 (Engineer, construction)
'LIBRARIAN I':                       2 (Librarian)
'LIBRARY PAGE':                      4 (Assistant (library))
'LIEUTENANT':                        1 (Lieutenant)
'LIEUTENANT-EMT':                    1 (Lieutenant)
'MACHINIST (AUTOMOTIVE)':            8 (Machinist (garage))
'MOTOR TRUCK DRIVER':                8 (Driver, truck)
'OPERATING ENGINEER-GROUP A':        2 (Engineer, operations (electricity supplier))
'OPERATING ENGINEER-GROUP C':        2 (Engineer, operations (electricity supplier))
'PARAMEDIC':                         3 (Paramedic)
'PARAMEDIC I/C':                     3 (Paramedic)
'PLUMBER':                           5 (Plumber)
'POLICE COMMUNICATIONS OPERATOR I':  4 (Assistant, administration (police service))
'POLICE COMMUNICATIONS OPERATOR II': 4 (Assistant, administration (police service))
'POLICE OFFICER':                    3 (Officer, police)
'POLICE OFFICER
 (ASSIGNED AS DETECTIVE)':           3 (Detective (police service))
'POLICE OFFICER
 (ASSIGNED AS EVIDENCE TECHNICIAN)': 3 (Officer, police)
'POLICE OFFICER / FLD TRNG OFFICER': 3 (Officer, police)
'POOL MOTOR TRUCK DRIVER':           8 (Driver, truck)
'SANITATION LABORER':                6 (Worker, healthcare (hospital service))
'SERGEANT':                          3 (Sergeant)
'TRAFFIC CONTROL AIDE-HOURLY':       7 (Assistant, control, traffic, air)
```

Figure 6.26: Correspondences between $Job\ Title$ values for the Chicago dataset and related major groups, together with index entries of reference.

- **FAIR-DB**: we conducted a 2-bin analysis with the usual \$90K threshold, keeping the same parameter values used in previous scenarios and applying the same criterion for the manual selection of the rules (target attribute in the RHS).

  ACFD Discovery detected 547 dependencies, whose number shrank to 122 when we filtered the results keeping only the rules containing the target attribute and its value. The automatic filtering operations performed by FAIR-DB further reduced this number from 122 to 38.

The first automatic ACFD selection resulted in a reduction in the number of dependencies from 38 to 18, while the subsequent ACFD completion and further selection generated 68 dependencies (including the original 18), 31 of which with a difference above the threshold.

At the end, we got $N = 10$ out of 31 dependencies in the ACFD user selection phase. The chosen ACFDs, together with the final scores, are displayed in Figure 6.27.

As we can see, the algorithm detected the same dependencies of the standard Chicago case, with the introduction of 4 more rules, 2 of which related to job title 2 (professional occupations), already observed as problematic in the 'Glassdoor Method' analysis. The other dependencies, not related to each other, suggest that females employed in job title 8 (process, plant and machine operatives) tend to earn less than \$90K, and the same holds for males employed in job title 4 (administrative and secretarial occupations).

The cumulative support is obviously higher than the one of the standard Chicago scenario, because the same rules have been detected, together with 4 more dependencies involving other tuples. The increment however is not significantly impactful on the results (from 11.4% to 12% of the dataset marked as 'problematic').

- **Ranking Facts**: the generalization process performed on the *Job Title* attribute did not have an impact on the number of tuples of the dataset, and therefore we still had to rely on the notebook version. The impact on the heatmap was also minimal, and we avoid to report it here because no new significant correlations between attributes were highlighted. We used the same attributes and weights of the standard Chicago case for the ranking (*Job Title*, *Department*, *Status*, *Salary or Hourly*, *Annual Salary*), and the results are summarized as follows:

  - **Recipe** and **Ingredients**: no particularly useful information from these widgets. As done before, we just report that the importance of each attribute used for the ranking is effectively equal to 1.
  - **Stability**: as for the standard Chicago case, the ranking resulted to be unstable both at top-10 (stability at 0.05) and overall (stability at 0.0).
  - **Fairness**: the statistical measures adopted provided us with the following results:

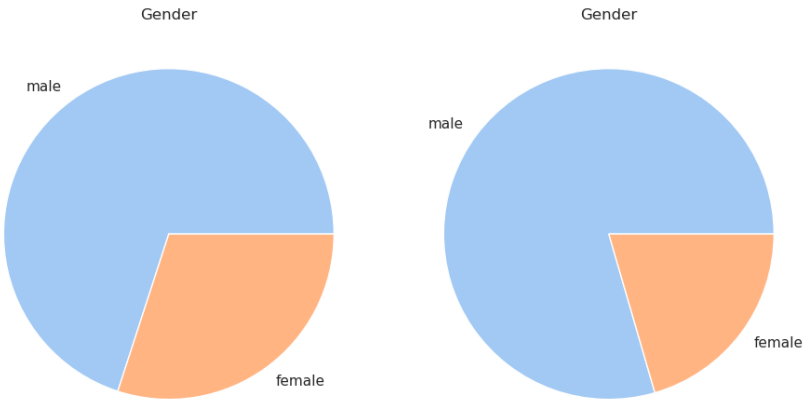
```
Number of tuples interested by the rules:  2442
Total number of tuples:  20309

Cumulative Support:  0.120
Difference Mean:  0.148
Gender - Difference Mean:  0.148


Total number of ACFDs selected:  10
                                        Rule  Supp  Conf  Diff  GenderDiff  Mean
47  {'lhs': {'Gender': 'female',
             'Job Title': '2'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  0.89  0.62        0.62  0.31
7   {'lhs': {'Gender': 'female',
             'Salary or Hourly': 'Hourly'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.03  0.89  0.24        0.24  0.14
44  {'lhs': {'Gender': 'male',
             'Job Title': '2'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.02  0.88  0.15        0.15  0.08
27  {'lhs': {'Gender': 'female',
             'Department': 'AVIATION'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.01  0.92  0.13        0.13  0.07
4   {'lhs': {'Gender': 'male',
             'Salary or Hourly': 'Hourly'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.07  0.41  0.06        0.06  0.06
39  {'lhs': {'Gender': 'female',
             'Job Title': '8'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.01  0.99  0.08        0.08  0.05
29  {'lhs': {'Gender': 'male',
             'Department': 'OEMC'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.01  0.94  0.08        0.08  0.05
41  {'lhs': {'Gender': 'male',
             'Job Title': '4'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.01  0.93  0.06        0.06  0.03
24  {'lhs': {'Gender': 'male',
             'Department': 'AVIATION'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.01  0.23  0.03        0.03  0.02
30  {'lhs': {'Gender': 'female',
             'Department': 'OEMC'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.00  0.18  0.03        0.03  0.02
```

Figure 6.27: Final selected rules and scores for the Chicago dataset with grouped job titles (2 bins).

- **FA*IR**: the ranking resulted to be *unfair for males*, with an approximate p-value of 0.03, and *fair for females*, with an approximate p-value of 1.0.
- **Proportion**: the ranking resulted to be *unfair for males*, with an approximate p-value of 0.05, and *fair for females*, with an approximate p-value of 0.99.



(a) *(b)*

Figure 6.28: Gender diversity widget for the top-10 (a) and overall (b) rankings of the Chicago dataset with grouped job titles.

  - ∗ **Pairwise**: the ranking resulted to be *unfair for males*, with an approximate p-value of 0.01, and *fair for females*, with an approximate p-value of 0.97.

  In contrast to the standard Chicago case and the outcomes of the other tools, the results seem to be oriented towards an unfair dataset, biased in favor of women.

- **Diversity**: Figure 6.28 shows the *Gender* diversity widget for the top-10 and overall rankings. As we can notice, even though there is still a predominance of the male group over the female one, in the top-10 ranking women seem to be slightly more represented than in the standard Chicago case.

6.3.5 Voluntary Introduction of Bias

In order to test the tools on a clearly biased dataset, we decided to modify the Chicago one by halving the *Annual Salary* value of female employees. This kind of operations generates a **synthetic dataset**, that is, a dataset containing fake data, not reflecting the real world in which we live. In other words, from the tools perspective, modifying the dataset means modifying the reality, and even if this consideration is applicable to every operation performed on data, in this specific context the original dataset undergoes such an impactful change that it becomes, as mentioned before, synthetic. Recalling the concepts discussed in Chapter 2, we can say we voluntarily

Dependent Variable: Log Annual Salary		
	(1)	(2)
Male	0.935	0.697
Job Title		0.115
Department		0.726
Status		0.362
Constant	10.461	10.377
Controls		
- Job Title	No	Yes
- Department	No	Yes
- Status	No	Yes
Observations	20309	20309
R^2	0.458	0.971

Figure 6.29: Regression results for the biased Chicago dataset.

introduced some technical bias with the aim of simulating the presence of preexisting bias in the society. Our main reason for generating a synthetic dataset is, in fact, to test the tools on data that we know for sure to be discriminatory.

Apart for the change in wage values of female employees, the preprocessed dataset is the same used in the standard Chicago case, with 20,309 tuples of which 16,146 males and 4,163 females. We will now provide the results of the analysis of the modified dataset:

- **The 'Glassdoor Method'**: for the sake of brevity, we avoid reporting summary table, pivot table, and average salaries for men and women, since the information provided by them reflects the dramatic modification on the income of female employees.

  Figure 6.29 shows instead our results: a coefficient of 0.935 on the male-female dummy variable means there is approximately 93.5% 'unadjusted' pay gap (therefore, men on average earn 93.5% more than women), and adding to the model all of the controls available in the data the coefficient value shrinks to 69.7%, remaining statistically significant. As we expected, the final outcome is significantly different from the standard Chicago case, and there is clear evidence of a systematic gender pay gap even on an 'adjusted' basis.

- **FAIR-DB**: we conducted a 2-bin analysis with the usual \$90K threshold, keeping the same parameter values used in previous scenarios and applying the same criterion for the manual selection of the rules (target

attribute in the RHS).

ACFD Discovery detected 785 dependencies, whose number shrank to 161 when we filtered the results keeping only the rules containing the target attribute and its value. The automatic filtering operations performed by FAIR-DB further reduced this number from 161 to 35.

The first automatic ACFD selection resulted in a reduction in the number of dependencies from 35 to 20, while the subsequent ACFD completion and further selection generated 79 dependencies (including the original 20), 39 of which with a difference above the threshold.

At the end, we got $N = 32$ out of 39 dependencies in the ACFD user selection phase. The chosen ACFDs, together with the final scores, are displayed in Figure 6.30. For readability reasons, the indication of the number of tuples concerned by the rules (13,311) and the total number of tuples (20,309) is not included in the image.

The algorithm detected pairs of dependencies related to 14 distinct *Job Title* values (out of 35), for which female employees earn less than \$90K while male employees earn more than \$90K. All of these are job titles in which the standard income is higher than the threshold, and this is the reason why the voluntary introduction of bias had an impact on the results. For the 21 other *Job Title* values, indeed, the average salary is lower than \$90K even for males, and the gender pay gap could not be observed. Decreasing the threshold value would be an option to include more job titles in the results, but then the opposite risk may occur: jobs for which the average income would be higher than the threshold even for females would not be detected. It is worth to mention the presence of rules 14 and 11, concerning the DAIS department, which has to be related one to one to a job title for which the average wage is higher than \$90K, and it is important to highlight the presence of rules 3 and 0, which 'generalize' the gender pay gap problem specifying that in general, regardless of the job title, women always earn less than \$90K while most men do not.

For what concerns the metrics, we can notice that for all the female-related rules (and for most of the others) the confidence parameter is equal to 1, meaning that these dependencies hold for all the tuples with the attribute values specified in the LHS. The cumulative support is high (65.5% of the dataset is 'problematic') because, even if most of the rules are related to specific job titles, rules 3 and 0 involve a significant amount of tuples. We can finally notice high values for the

```
Cumulative Support:  0.655
Difference Mean:  0.475
Gender - Difference Mean:  0.475


Total number of ACFDs selected:  32
                                                Rule  Supp  Conf  Diff  GenderDiff  Mean
62  {'lhs': {'Gender': 'female',
             'Job Title': 'HOISTING ENGINEER'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  1.00  0.98        0.98  0.49
50  {'lhs': {'Gender': 'female',
             'Job Title': 'CAPTAIN-EMT'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  1.00  0.98        0.98  0.49
46  {'lhs': {'Gender': 'female',
             'Job Title': 'LIEUTENANT-EMT'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  1.00  0.97        0.97  0.49
58  {'lhs': {'Gender': 'female',
             'Job Title': 'PLUMBER'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  1.00  0.97        0.97  0.48
26  {'lhs': {'Gender': 'female',
             'Job Title': 'FIRE ENGINEER-EMT'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  1.00  0.96        0.96  0.48
42  {'lhs': {'Gender': 'female',
             'Job Title': 'OPERATING ENGINEER-GROUP C'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  1.00  0.95        0.95  0.48
34  {'lhs': {'Gender': 'female',
             'Job Title': 'FIREFIGHTER-EMT'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  1.00  0.95        0.95  0.48
22  {'lhs': {'Gender': 'female',
             'Job Title': 'ELECTRICAL MECHANIC'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  1.00  0.94        0.94  0.47
14  {'lhs': {'Gender': 'female',
             'Department': 'DAIS'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  1.00  0.92        0.92  0.46
66  {'lhs': {'Gender': 'female',
             'Job Title': 'LIEUTENANT'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  1.00  0.88        0.88  0.44
54  {'lhs': {'Gender': 'female',
             'Job Title': 'CONSTRUCTION LABORER'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  1.00  0.86        0.86  0.43
70  {'lhs': {'Job Title': 'SERGEANT',
             'Gender': 'female'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.01  1.00  0.84        0.84  0.42
30  {'lhs': {'Gender': 'female',
             'Job Title': 'FIREFIGHTER/PARAMEDIC'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  1.00  0.78        0.78  0.39
18  {'lhs': {'Gender': 'female',
             'Department': 'WATER MGMNT'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  1.00  0.75        0.75  0.38
38  {'lhs': {'Gender': 'female',
             'Job Title': 'PARAMEDIC I/C'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.00  1.00  0.71        0.71  0.36
```

```
3   {'lhs': {'Gender': 'female'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.20  1.00  0.45        0.45  0.33
0   {'lhs': {'Gender': 'male'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.45  0.57  0.12        0.12  0.28
35  {'lhs': {'Gender': 'male',
              'Job Title': 'PARAMEDIC I/C'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.01  1.00  0.29        0.29  0.15
67  {'lhs': {'Job Title': 'SERGEANT',
              'Gender': 'male'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.05  1.00  0.16        0.16  0.10
51  {'lhs': {'Gender': 'male',
              'Job Title': 'CONSTRUCTION LABORER'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.02  1.00  0.14        0.14  0.08
63  {'lhs': {'Gender': 'male',
              'Job Title': 'LIEUTENANT'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.01  1.00  0.12        0.12  0.06
31  {'lhs': {'Gender': 'male',
              'Job Title': 'FIREFIGHTER-EMT'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.06  1.00  0.05        0.05  0.06
15  {'lhs': {'Gender': 'male',
              'Department': 'WATER MGMNT'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.04  0.82  0.07        0.07  0.05
27  {'lhs': {'Gender': 'male',
              'Job Title': 'FIREFIGHTER/PARAMEDIC'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.01  0.85  0.08        0.08  0.04
19  {'lhs': {'Gender': 'male',
              'Job Title': 'ELECTRICAL MECHANIC'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.01  1.00  0.06        0.06  0.03
23  {'lhs': {'Gender': 'male',
              'Job Title': 'FIRE ENGINEER-EMT'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.02  1.00  0.04        0.04  0.03
39  {'lhs': {'Gender': 'male',
              'Job Title': 'OPERATING ENGINEER-GROUP C'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.01  1.00  0.05        0.05  0.03
43  {'lhs': {'Gender': 'male',
              'Job Title': 'LIEUTENANT-EMT'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.02  1.00  0.03        0.03  0.02
11  {'lhs': {'Gender': 'male',
              'Department': 'DAIS'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.01  0.94  0.03        0.03  0.02
55  {'lhs': {'Gender': 'male',
              'Job Title': 'PLUMBER'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.01  1.00  0.03        0.03  0.02
59  {'lhs': {'Gender': 'male',
              'Job Title': 'HOISTING ENGINEER'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.01  1.00  0.02        0.02  0.02
47  {'lhs': {'Gender': 'male',
              'Job Title': 'CAPTAIN-EMT'},
     'rhs': {'Annual Salary Bin': '> 90K'}}   0.01  1.00  0.02        0.02  0.02
```

Figure 6.30: Final selected rules and scores for the biased Chicago dataset (2 bins).

difference measure for all the female-related rules, indicating a high 'unethical' level towards women.

- **Ranking Facts**: the voluntary introduction of bias did not have an impact on the number of tuples of the dataset, and therefore we still had to rely on the notebook version. The impact on the heatmap was also minimal, and we avoid to report it here because no new significant correlations between attributes were highlighted. We used the same attributes and weights of the standard Chicago case for the ranking (*Job Title*, *Department*, *Status*, *Salary or Hourly*, *Annual Salary*), and the results are summarized as follows:

  - **Recipe** and **Ingredients**: no particularly useful information from these widgets. As done before, we just report that the importance of each attribute used for the ranking is effectively equal to 1.
  - **Stability**: as for the standard Chicago case, the ranking resulted to be unstable both at top-10 (stability at 0.07) and overall (stability at 0.0).
  - **Fairness**: the statistical measures adopted provided us with the following results:
    - **FA*IR**: the ranking resulted to be *fair for males*, with an approximate p-value of 1.0, and *unfair for females*, with an approximate p-value of 0.0.
    - **Proportion**: the ranking resulted to be *fair for males*, with an approximate p-value of 1.0, and *unfair for females*, with an approximate p-value of 0.0.
    - **Pairwise**: the ranking resulted to be *fair for males*, with an approximate p-value of 1.0, and *unfair for females*, with an approximate p-value of 0.0.

    The voluntary introduction of bias resulted, for each metric, in a change in the female-related outcome in comparison with the standard Chicago case. As we expected, the results seem to be oriented towards an unfair dataset, in which women are discriminated against.
  - **Diversity**: Figure 6.31 shows the *Gender* diversity widget for the top-10 and overall rankings. As we expected, the representation problem is further accentuated, with the top-10 ranking monopolized by men.

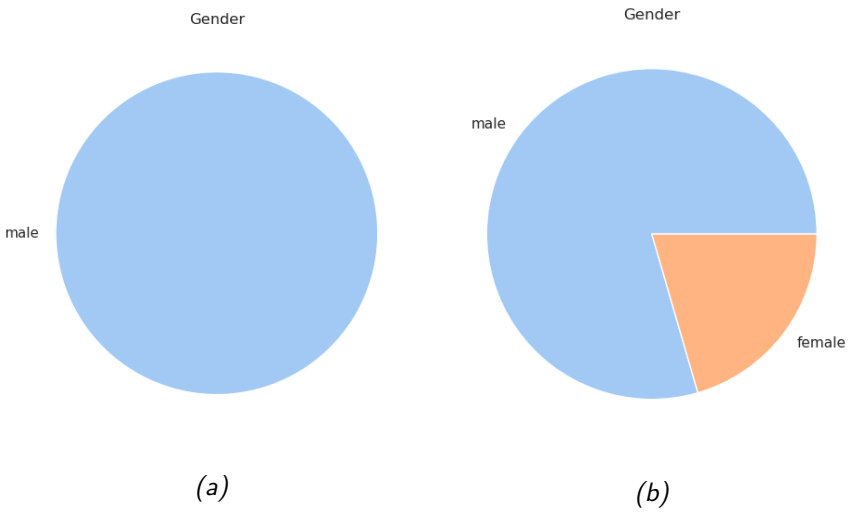


Figure 6.31: Gender diversity widget for the top-10 (a) and overall (b) rankings of the biased Chicago dataset.

Lastly, we decided to examine the behavior of the tools on two other synthetic datasets, generated from the standard Chicago one, in which we retained respectively 75% and 90% of the *Annual Salary* value of female employees, generating in fact two datasets 'a little less biased' than the case just analyzed.

Again, apart for the change in wage values of female employees, the preprocessed datasets are the same used in the standard Chicago case, with 20,309 tuples of which 16,146 males and 4,163 females. Without going too much into detail, we will now provide the results of the analysis of the modified datasets:

- **The 'Glassdoor Method'**: the algorithm detected an 'adjusted' pay gap of 29.1% and 10.9% respectively, in line with our predictions. Indeed, we were expecting a more or less linear progression from the halved salary case (69.7%) to the standard case (0.4%). Both the results remain statistically significant.

- **FAIR-DB**: we conducted a 2-bin analysis with the usual \$90K threshold, keeping the same parameter values used in previous scenarios and applying the same criterion for the manual selection of the rules (target attribute in the RHS). At the end, we got respectively 43 and 22 dependencies.

  The fact that the algorithm detected more dependencies in the 75% scenario than in the 50% one (in which we got 32) may seem a little

surprising, but by looking more closely at the rules generated we found the reasons for this behavior.

First of all, most of the dependencies are the same as the halved salary case, showed in Figure 6.30, with the exception of rules 50, 46, 66, 63, 43, and 47, which do not appear in the 75% case. On the other hand, in addition to the preexisting 26 rules (the 32 of the 50% case minus the 6 just mentioned), also the ones displayed in Figure 6.32 are generated, bringing the total to 43.

The 6 rules of the halved salary case which do not show up here are related to the CAPTAIN-EMT, LIEUTENANT-EMT, and LIEUTENANT job titles, because for those jobs women are paid less than the threshold, while men are not. In the 75% case, these rules are no longer generally valid, because 163 women employed in these professions earn more than $90K. Therefore, the algorithm detects rules related to the 'side attributes' of those tuples, because for them women generally earn less than $90K and men earn more. Specifically, even if for example rules 50 and 47 of Figure 6.30 (related to the CAPTAIN-EMT profession) are no longer valid, because many women employed in this job earn more than the threshold, this is generally not true for other employees in the FIRE department (of which captains emeritus are also part), and consequently rules 43 and 40 of Figure 6.32 are generated; and it is generally not true even with regard to full-time employees (rules 31 and 28), and employees paid on a salary basis (rules 15 and 12). The same holds for the POLICE department (rules 59 and 56) and for the POLICE OFFICER (rules 91 and 88) and POLICE COMMUNICATIONS OPERATOR II (rules 84 and 87) job titles, connected to the POLICE department.

Furthermore, since some women earn more than $90K, the confidence of rule 3 of Figure 6.30 decreases from 1.0 to 0.96, and therefore the rule does not involve automatically the whole group of female employees. For this reason, the algorithm detects further dependencies related to employees whose pay frequency is Hourly (rules 11 and 8) and departments for which the majority (or even the totality) of the employees are paid on an hourly basis, namely AVIATION (rules 47 and 44) and STREETS & SAN (rule 39).

For what concerns the 90% scenario, as we expected, the 22 dependencies obtained are a subset of the ones we got from the 50% and 75% cases, and therefore we avoid to report them again. Instead, it is important to highlight that the cumulative support decreases coherently

	Rule	Supp	Conf	Diff	GenderDiff	Mean
43	{'lhs': {'Gender': 'female', 'Department': 'FIRE'}, 'rhs': {'Annual Salary Bin': '<= 90K'}}	0.02	0.96	0.69	0.69	0.35
31	{'lhs': {'Status': 'F', 'Gender': 'female'}, 'rhs': {'Annual Salary Bin': '<= 90K'}}	0.18	0.96	0.43	0.43	0.30
15	{'lhs': {'Salary or Hourly': 'Salary', 'Gender': 'female'}, 'rhs': {'Annual Salary Bin': '<= 90K'}}	0.16	0.95	0.44	0.44	0.30
28	{'lhs': {'Status': 'F', 'Gender': 'male'}, 'rhs': {'Annual Salary Bin': '> 90K'}}	0.45	0.58	0.10	0.10	0.28
12	{'lhs': {'Salary or Hourly': 'Salary', 'Gender': 'male'}, 'rhs': {'Annual Salary Bin': '> 90K'}}	0.38	0.61	0.12	0.12	0.25
59	{'lhs': {'Department': 'POLICE', 'Gender': 'female'}, 'rhs': {'Annual Salary Bin': '<= 90K'}}	0.12	0.94	0.38	0.38	0.25
91	{'lhs': {'Gender': 'female', 'Job Title': 'POLICE OFFICER'}, 'rhs': {'Annual Salary Bin': '<= 90K'}}	0.10	1.00	0.31	0.31	0.21
11	{'lhs': {'Gender': 'female', 'Salary or Hourly': 'Hourly'}, 'rhs': {'Annual Salary Bin': '<= 90K'}}	0.04	1.00	0.33	0.33	0.18
56	{'lhs': {'Department': 'POLICE', 'Gender': 'male'}, 'rhs': {'Annual Salary Bin': '> 90K'}}	0.24	0.55	0.11	0.11	0.17
88	{'lhs': {'Gender': 'male', 'Job Title': 'POLICE OFFICER'}, 'rhs': {'Annual Salary Bin': '> 90K'}}	0.13	0.41	0.10	0.10	0.11
40	{'lhs': {'Gender': 'male', 'Department': 'FIRE'}, 'rhs': {'Annual Salary Bin': '> 90K'}}	0.14	0.79	0.06	0.06	0.10
47	{'lhs': {'Gender': 'female', 'Department': 'AVIATION'}, 'rhs': {'Annual Salary Bin': '<= 90K'}}	0.01	1.00	0.19	0.19	0.10
84	{'lhs': {'Gender': 'male', 'Job Title': 'POLICE COMMUNICATIONS OPERATOR II'}, 'rhs': {'Annual Salary Bin': '> 90K'}}	0.00	0.23	0.19	0.19	0.09
8	{'lhs': {'Gender': 'male', 'Salary or Hourly': 'Hourly'}, 'rhs': {'Annual Salary Bin': '> 90K'}}	0.07	0.41	0.08	0.08	0.07
87	{'lhs': {'Gender': 'female', 'Job Title': 'POLICE COMMUNICATIONS OPERATOR II'}, 'rhs': {'Annual Salary Bin': '<= 90K'}}	0.01	1.00	0.04	0.04	0.03
44	{'lhs': {'Gender': 'male', 'Department': 'AVIATION'}, 'rhs': {'Annual Salary Bin': '> 90K'}}	0.01	0.23	0.04	0.04	0.02

```
39  {'lhs': {'Gender': 'female',
             'Department': 'STREETS & SAN'},
     'rhs': {'Annual Salary Bin': '<= 90K'}}  0.01  1.00  0.02        0.02  0.02
```

Figure 6.32: Additional final selected rules and scores for the biased Chicago dataset with 75% of the Annual Salary value of female employees (2 bins).

with the reduction of the *Annual Salary* value of female employees, going from 0.655 (50% case) to 0.647 (75% case), then to 0.629 (90% case) and finally to 0.114 (standard case). As we can notice, even if the reduction of the *Annual Salary* value of female employees is not particularly dramatic ('just' 10% less than the standard case), the number of tuples involved is very significant, which is to the advantage of FAIR-DB, thus making it able to detect even gaps that are not incredibly wide.

- **Ranking Facts**: the tool did not provide us with different results than the halved salary case, neither from the point of view of fairness nor with regard to diversity. Furthermore, like in the previous case, the ranking was found to be unstable both at top-10 and overall for both the 75% case (top-10 stability at 0.07) and the 90% one (top-10 stability at 0.08).

Chapter 7

Conclusions & Future Work

The aim of this chapter is to draw the conclusions of our research, mixing the results obtained in our experiments described in Chapter 6 with the sociological background depicted in Chapter 4, and recalling some preliminaries exposed in Chapter 2 and Chapter 3 when needed.

We will start by recapitulating our work in a *conclusive summary*, and then we will discuss about the *outcomes and contributions* of our research, emphasizing also the *limitations* to which we were subjected. Finally, we will suggest *future work* to be potentially done starting from this research, and some possible paths that might be worth to follow.

7.1 Conclusive Summary

Starting from the general notions of fairness and data management, we reviewed both socio-ethical and technical literature in order to dive deep into the topics and capture the several facets of the problem of discovering bias in the data.

For what concerns the socio-ethical side, we classified three categories of bias, different in origin and in nature; we clarified what discrimination is, and how it is connected to human rights; and we distinguished between equality and equity, extending the concepts to the data perspective, and finally pointing at what both approaches aim to achieve: fairness.

On the technical side, we first introduced the basics of the discipline by providing explanations on relational databases, data science pipeline, and data mining techniques; we then explored some more specific concepts such as linear regression and functional dependencies, also providing detailed information on some evaluation metrics for them, and on some other useful statistical concepts; and we finally described, referring to the documentation

at our disposal, the tools we decided to adopt for our analysis: the 'Glassdoor Method', FAIR-DB, and Ranking Facts.

After this first literature review phase, we conducted parallel research in sociology and information technology.

Once chosen the datasets on which to work and the specific focus of our analysis – gender gap – we retrieved information on the problem in the society, overall at first, introducing The Global Gender Gap Index, and then focusing on the U.S., providing different point of views and reasons for the presence of gender discrimination in the workplace, such as statistical discrimination, impact of the institutional environment, and inequalities related to unequal bargaining power; and finally we showed some data and statistics from the U.S. Department of Labor, in order to make more concrete the reflections previously reported.

Meanwhile, we did some experiments on our datasets, in order to verify the possible presence of bias in the data, potentially leading to gender discrimination and unfair outcomes. For both our case studies, after having provided a more detailed description of the dataset, we performed some data preprocessing operations; and we used the tools previously described to analyze them. We tried to identify – both during the preprocessing and the actual analysis – the most critical choices we had to deal with, pointing them as potential sources of bias; and we conducted other experiments on the same data but taking different paths, ultimately evaluating the impact of different design decisions on our outcomes.

7.2 Outcomes & Contributions

Combining the sociological research with the technological results of our analysis, we can highlight two main issues: a *representation problem*, related to the disproportion in the percentage of women employed in different sectors, and a *part-time problem*, due to the higher number of women employed in part-time jobs, typically less paid than full-time ones.

For what concerns representation, we want to recall the data and statistics provided in Section 4.3 related to the most common occupations for women and the percentage of women employed in STEM disciplines, and point out that, even though there is clear evidence of the presence of this problem in our datasets, as we have mentioned throughout Chapter 6, none of the adopted tools highlights this particular aspect (Ranking Facts tries to encompass it through its diversity widget, with the result of providing just a high-level and absolutely non-exhaustive visual representation that cannot capture the underlying complexity and the different facets of the issue).

The part-time condition, for which we recall Figure 4.2 and, in general, the reflections made in Chapter 4, is instead more perceptible from the tools, since the information is encoded in our datasets under the *Status* attribute, and therefore could be used for estimating the 'adjusted' pay gap in the 'Glassdoor Method', as constituent of functional dependencies in FAIR-DB, and as a ranking parameter in Ranking Facts. Removing the part-time employees instead resulted to be penalizing both from the technical point of view, as underlined in Section 6.3.1, and from the sociological one, given the background provided in Chapter 4.

The most impactful limitation of the tools, in any case, is the fact that even though they all aim to achieve fairness not only through equality but also by taking equity into account (by privileging the group fairness criteria over the individual fairness one), they practically fail in capturing the several facets of equity. Recalling the classification made in Section 2.4, we proved how inefficient these instruments are in capturing *representation equity* (disparities rooted in historical discrimination can lead to representation inequities); *feature equity* (not all the features needed to represent a marginalized group and required for a particular analysis are available in the data) is not addressed at all, since the tools do not have knowledge on the context, and therefore they cannot, for example, infer what the relevant attributes for the specific analysis may be and inform the user on what is missing; and neither are addressed *access equity* (bias may arise because of a non-equitable and participatory access to data and data products across domains and levels of expertise) and *outcome equity* (bias may arise because of a lack of monitoring and mitigation of unintended consequences for any group affected by the system after deployment).

Moving from outcomes to contributions: first of all we confirmed, in accordance with Chouldechova's impossibility theorem [13] and however trivial it may be, that fairness is a multifaceted concept which cannot be exhausted by providing a single definition and pursuing that specific definition experimentally. In this context, we believe that the simultaneous use of several tools for the same analysis represents an undoubted advantage, since each of them provides a different approach and hence a different perspective on the same problem.

We also demonstrated how these 'fairness measurement tools' are susceptible to decisional choices, and therefore how important it is to properly train users on the specific area of analysis. A question that might arise is: 'Since human intervention (which takes the form of choices in the use of these tools) necessarily introduces, being human, bias, does it make sense to keep humans in the loop?'. On the other hand, we could ask ourselves: 'Given the problems

involved in developing a universal instrument that can address the issue in all its facets, does it make sense to manage decision-making processes with these tools?'. We claim the answer to be yes in both cases: fairness indeed is a complex human concept of abstract nature, with impactful concrete social implications, and as such it requires a non-automatable human intervention; on the other hand it is not possible to do without the tools – that it is necessary to continue to build and improve – since the amount of data to handle is absolutely excessive and not manageable otherwise.

Lastly, we believe that the main contribution of this research is given by the double perspective on the gender pay gap issue, which we hope will set a precedent for data scientists and the other professionals in the IT sector for approaching other social problems. We strongly believe in multidisciplinarity and in the potentials of facing challenges not only by looking at the knowledge available in one's field of study but also in other, possibly interconnected, disciplines.

7.3 Limitations

Although our research provided us with some significant results, there are of course aspects of our work which limit the impact or generalizability of our contributions.

From the sociological point of view, the most important limitation is given by the fact that the literature of reference is usually non-specific, and although the majority of our sources is related to the U.S. society, the United States is a very wide country which also presents significant differences between the various states that make it, and none of the papers specifically refers to Illinois or California, and neither goes into detail of the cities of Chicago or San Francisco. Furthermore, we had to rely on publicly available data and resources, and we did not get any insight from employees or employers working in the U.S., so our view on the overall situation may be partial.

From the technical perspective, we already mentioned throughout Chapter 6 some of the most impactful choices we had to deal with, which potentially lead to the introduction of bias, but we will now recapitulate providing a more exhaustive summary.

- The original datasets are already partial, because they contain a limited number of job titles, and we are not aware of possible preexisting groupings of employees who may work on some specific tasks but be grouped in the same category, or of external employees who just

temporarily work for the federal government.

- The `gender-guesser` package we used to infer employees' gender is obviously not 100% accurate, and even if we assumed mostly male names to be effectively related to males and mostly female names to be effectively related to females, we may have been mistaken in some cases. Furthermore, the package produced a non-negligible number of unknown and androgynous occurrences, possibly related to employees of various ethnicities with not typically Western names, and even if we decided to remove them in order to work on more reliable data, they may have had an impact on the results.

- The data cleaning processes we performed by removing job titles with less than 100 occurrences significantly reduced the number of tuples of our datasets, and even if this choice led to a reduction of complexity, the downside is that a lot of categories of employees were excluded from the analysis.

- As previously pointed out, none of the adopted tools takes into account overrepresentation of men (or underrepresentation of women) in the specific job title in which they are employed, and although we can see it is as an outcome for the purpose of our research, it is certainly also a limitation, being a hidden bias source.

- The parameter values required for the FAIR-DB analysis have an impact on the number of dependencies detected and selected, and even if we used what we believe to be the best values, a deeper knowledge of the underlying concepts may have led to a refinement of the results. The same holds for the manual selection of the rules, that is a crucial phase of the framework, with the power to overturn the final results and for which, given the lack of documentation, we adopted what we think is the most sensible and suitable criterion. Choosing instead all the detected rules, as shown in Section 6.3.3, strongly impact the final outcomes, reversing the perspective on the fairness of data.

- Given the lack of documentation for Ranking Facts, we followed the approach of the authors of the tool by trying to laboriously follow the examples they provided in setting up the ranking parameters and their weights, but again a deeper knowledge of the underlying concepts may have led to a refinement of the results. Furthermore, given the size of our datasets, we were forced to use the notebook version of

the tool, which is undoubtedly less user-friendly than the Web-based application.

- The number of bins used for the FAIR-DB analysis and the values specified for the thresholds have an impact on the results. We tried to overcome this limit by conducting two different analyses, with respectively 2 and 8 bins (Section 6.3.2), but other choices may have led to different outcomes.

- As already mentioned in Section 6.3.4, for grouping job titles we relied on a document published by the Equality Commission for Northern Ireland in 2013 [23]. Despite the exhaustiveness of the index, it is important to underline that the document was not made with the purpose of being used for grouping job titles outside Northern Ireland.

- As already mentioned in Section 6.3.5, the voluntary introduction of bias in our data produced a synthetic dataset, that is, a dataset containing fake data, not reflecting the real world in which we live.

7.4 Future Work

Assuming we have more time (and maybe more resources) at our disposal, here we want to highlight some aspects which we think would deserve further study or development in future research.

First of all, having seen the different approaches of the adopted tools and the different perspectives they provide, it would be interesting to combine them all in a unique, more complete instrument, in order to give the user a single tool that provides multiple points of view, rather than several partial ones which the user themselves may not combine, maybe because not aware of the existence of each. Further efforts may be invested in trying to encompass even more facets of equity, or more definitions of fairness.

Even assuming such an instrument is developed, however, analyses of these kind should always be supported by sociological research, in order to get a broader perspective on the problem and capture facets which would not be captured otherwise (in our case, the representation issue).

The sociological research could also be further enriched by conducting an interview with workers and HR practitioners of the cities under study, in order to get more specific, recent, and precise information useful for interpreting the results.

For what concerns datasets, it would be appropriate to retrieve further information in support of the mere data, in order to get a more exhaustive

overview, and since since having more information usually leads to more accurate results. This also could be an incentive for developers, database administrators, and other professionals in the IT sector, to create effective documentation in support of data and technological tools. This documentation should, as far as possible, be detailed and at the same time easy to experience, so as to potentially enable professionals from other sectors (for example, sociologists) to get an idea of what is included in the data or how to use a tool in a conscious way. In this regard, we think that context-awareness would be an interesting path to follow, and some techniques may be used to provide the tool(s) with knowledge on the context of use. We believe that such an improvement would mitigate (even if not extinguish) problems, like the representation one, encapsulated in the data but not currently detectable. It is worth reporting here a couple of contributions that we think may be good starting points for addressing data quality issues. In particular, in [10], Canali argues that:

> *Quality is a contextual feature of data: it is a result of the relations established between a dataset and the questions, aims and tools employed in the context of the use of data; the assessment of the quality of a dataset needs to focus on the features of this context as much as the dataset itself.* [10, p. 4]

He then delineates some guidelines according to which the contextual approach indicates quality criteria and assessment methods, and provides three practical examples in support of it. In [36] instead, Leonelli examines some models of data quality evaluation that have been employed within the sciences, highlighting strengths and weaknesses of each and ultimately emphasizing again the importance of the context and of having exhaustive metadata in support of the mere data, in order also to facilitate international cooperation and the development of different projects on the same, accessible, data.

References

[1] Social Justice - Overview, History and Evolution, Five Principles. *Corporate Finance Institute*, 2020. https://corporatefinanceinstitute.com. Accessed 8 June 2021.

[2] Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*, volume 8. Addison-Wesley Reading, 1995. http://webdam.inria.fr/Alice.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias. *ProPublica*, 2016. https://www.propublica.org.

[4] UN General Assembly et al. Universal Declaration of Human Rights. *UN General Assembly*, 302(2):14–25, 1948. doi:https://doi.org/10.1017/9781316677117.029.

[5] Fabio Azzalini, Chiara Criscuolo, and Letizia Tanca. FAIR-DB: FunctionAl DependencIes to discoveR Data Bias. In *EDBT/ICDT Workshops*, 2021. http://ceur-ws.org/Vol-2841.

[6] Robert L Barker et al. *The Social Work Dictionary*. NASW Press, Washington, DC, 2003.

[7] John J Beggs. The Institutional Environment: Implications for Race and Gender Inequality in the U.S. Labor Market. *American Sociological Review*, pages 612–633, 1995. doi:10.2307/2096297.

[8] The Editors of Encyclopaedia Britannica. "Equality". *Encyclopedia Britannica*, 2009. https://www.britannica.com. Accessed 7 June 2021.

[9] The Editors of Encyclopaedia Britannica. "Database". *Encyclopedia Britannica*, 2020. https://www.britannica.com. Accessed 10 June 2021.

[10] Stefano Canali. Towards a Contextual Approach to Data Quality. *Data*, 5(4):90, 2020. doi:10.3390/data5040090.

[11] Loredana Caruccio, Vincenzo Deufemia, and Giuseppe Polese. Relaxed Functional Dependencies–A Survey of Approaches. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):147–165, 2015. doi:10.1109/tkde.2015.2472010.

[12] Andrew Chamberlain. How to Analyze Your Gender Pay Gap: An Employer's Guide. *Glassdoor Economic Research*, 2017. https://www.glassdoor.com.

[13] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017. doi:10.1089/big.2016.0047.

[14] Edgar F Codd. A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM*, 13(6):377–387, 1970. doi:10.1145/362384.362685.

[15] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, 2018. https://www.reuters.com.

[16] Cambridge Advanced Learner's Dictionary. "Data". *Cambridge University Press*, 2013. https://dictionary.cambridge.org. Accessed 10 June 2021.

[17] Cambridge Advanced Learner's Dictionary. "Fairness". *Cambridge University Press*, 2013. https://dictionary.cambridge.org. Accessed 15 June 2021.

[18] Cambridge Advanced Learner's Dictionary. "Gender gap". *Cambridge University Press*, 2013. https://dictionary.cambridge.org. Accessed 28 June 2021.

[19] Selin Dilli, Sarah G Carmichael, and Auke Rijpma. Introducing the Historical Gender Equality Index. *Feminist Economics*, 25(1):31–57, 2018. doi:10.1080/13545701.2018.1442582.

[20] Paul J DiMaggio and Walter W Powell. The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields. *American Sociological Review*, pages 147–160, 1983. doi:doi:10.2307/2095101.

[21] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012. arXiv:1104.3913.

[22] Nancy Folbre. Gender inequality and bargaining in the U.S. labor market. *Economic Policy Institute*, 2021. https://www.epi.org/209716.

[23] Equality Commission for Northern Ireland. *Index for Classifying Job Titles*. Equality Commission for Northern Ireland, Equality House 7–9, Shaftesbury Square, Belfast BT2 7DP, 2013. https://www.equalityni.org.

[24] World Economic Forum. Terms of Use. https://www.weforum.org. Accessed 28 July 2021.

[25] Batya Friedman and Helen Nissenbaum. Bias in Computer Systems. In *Computer Ethics*, pages 215–232. Routledge, 2017. doi:10.4324/9781315259697-23.

[26] Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144, 2015. doi:10.1016/j.ijinfomgt.2014.10.007.

[27] Samuel Gibbs. Women less likely to be shown ads for high-paid jobs on Google, study shows. *The Guardian*, 8(7), 2015. https://www.theguardian.com.

[28] Paul Glasserman. Linear Regression. *Lecture notes in Managerial Statistics*, 2001. https://www.gsb.columbia.edu.

[29] Kelly L Hazel and Kerry S Kleyman. Gender and sex inequalities: Implications and resistance, 2019. doi:10.1080/10852352.2019.1627079.

[30] Jeff Horwitz. Facebook Algorithm Shows Gender Bias in Job Ads, Study Finds. *The Wall Street Journal*, 2021. https://www.wsj.com.

[31] HV Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, and Cyrus Shahabi. Big Data and Its Technical Challenges. *Communications of the ACM*, 57(7):86–94, 2014. doi:10.1145/2611567.

[32] HV Jagadish, Julia Stoyanovich, and Bill Howe. COVID-19 Brings Data Equity Challenges to the Fore. *Digital Government: Research and Practice*, 2(2):1–7, 2021. doi:10.1145/3440889.

[33] HV Jagadish, Julia Stoyanovich, and Bill Howe. The Many Facets of Data Equity. In *24th International Conference on Extending Database Technology (EDBT)*, 2021. http://ceur-ws.org/Vol-2841.

[34] Alexandros Labrinidis and HV Jagadish. Challenges and Opportunities with Big Data. *Proceedings of the VLDB Endowment*, 5(12):2032–2033, 2012. doi:10.14778/2367502.2367572.

[35] Heidi Ledford. Millions of black people affected by racial bias in healthcare algorithms. *Nature*, 574(7780):608–610, 2019. doi:10.1038/d41586-019-03228-6.

[36] Sabina Leonelli. Global Data Quality Assessment and the Situated Nature of "Best" Research Practices in Biology. *Data Science Journal*, 16, 2017. doi:10.5334/dsj-2017-032.

[37] G Marschall. *A Dictionary of Sociology*. Oxford University Press, 1998.

[38] Andy Mason. "Equal opportunity". *Encyclopedia Britannica*, 2019. https://www.britannica.com. Accessed 7 June 2021.

[39] Steve Mills, Steve Lucas, Leo Irakliotis, Michael Rappa, Teresa Carlson, and Bill Perlowitz. Demystifying Big Data: A Practical Guide To Transforming The Business of Government. *TechAmerica Foundation, Washington*, 2012. https://bigdatawg.nist.gov.

[40] Jonathon Penney, Sarah McKune, Lex Gill, and Ronald J Deibert. Advancing Human-Rights-by-Design in the Dual-Use Technology Industry. *Journal of International Affairs*, 71(2):103–110, 2018. https://www.jstor.org/stable/10.2307/26552332.

[41] Joeri Rammelaere and Floris Geerts. Revisiting Conditional Functional Dependency Discovery: Splitting the "C" from the "FD". In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 552–568. Springer, 2018. doi:10.1007/978-3-030-10928-8_33.

[42] Thomas L Saaty and Mujgan S Ozdemir. Why the Magic Number Seven Plus or Minus Two. *Mathematical and Computer Modelling*, 38(3-4):233–244, 2003. doi:10.1016/s0895-7177(03)90083-5.

[43] Teresa Scantamburlo, Andrew Charlesworth, and Nello Cristianini. Machine decisions and human consequences. In Karen Yeung and Martin Lodge, editors, *Algorithmic Regulation*, pages 49–81. Oxford University Press, 2019. doi:10.1093/oso/9780198838494.003.0003.

[44] Klaus Schwab, Richard Samans, Saadia Zahidi, Till Alexander Leopold, Vesselina Ratcheva, Ricardo Hausmann, and Laura D Tyson. The Global Gender Gap Report 2017. World Economic Forum, 2017. https://www.weforum.org.

[45] M Sepuldeva, Th Van Banning, Gudrún Gudmundsdóttir, Christine Chamoun, and Willem JM Van Genugten. *Human Rights Reference Handbook*. University for Peace, 2010.

[46] R Tamilselvi, B Sivasakthi, and R Kavitha. An efficient preprocessing and postprocessing techniques in data mining. *International Journal of Research in Computer Applications and Robotics*, 3(4):80–85, 2015. https://www.ijrcar.com.

[47] András Tilcsik. Statistical Discrimination and the Rationalization of Stereotypes. *American Sociological Review*, 2021. doi:10.31235/osf.io/gf7ry.

[48] Sahil Verma and Julia Rubin. Fairness Definitions Explained. In *2018 2018 ACM/IEEE International Workshop on Software Fairness*, pages 1–7. IEEE, 2018. doi:10.1145/3194770.3194776.

[49] Yongkai Wu, Lu Zhang, and Xintao Wu. Counterfactual Fairness: Unidentification, Bound and Algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 1438–1444. International Joint Conferences on Artificial Intelligence Organization, 2019. doi:10.24963/ijcai.2019/199.

[50] Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, HV Jagadish, and Gerome Miklau. A Nutritional Label for Rankings. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1773–1776. Association for Computing Machinery, 2018. doi:10.1145/3183713.3193568.

[51] Karen Yeung, Andrew Howes, and Ganna Pogrebna. AI Governance by Human Rights-Centered Design, Deliberation and Oversight: An End to Ethics Washing. *The Oxford Handbook of Ethics of AI*, page 77, 2020. doi:10.1093/oxfordhb/9780190067397.013.5.

[52] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1569–1578. Association for Computing Machinery, 2017. doi:10.1145/3132847.3132938.

[53] Indrė Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089, 2017. doi:10.1007/s10618-017-0506-1.

Appendix A

Country Profile of the United States (The Global Gender Gap Report 2017)

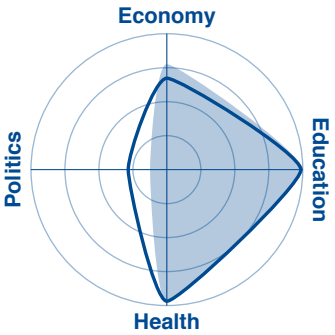
United States

rank 49
out of 144 countries

score 0.718
0.00 = imparity
1.00 = parity

0.40 distribution of countries by score 1.00

SCORE AT GLANCE



United States score
average score

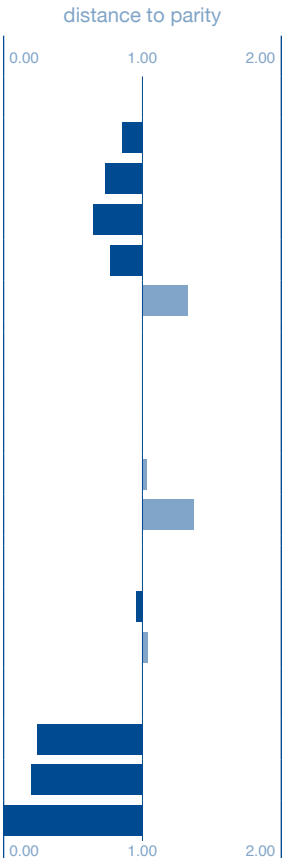
KEY INDICATORS

GDP (US$ billions)	18,569.10
GDP per capita (constant '11, intl. $, PPP)	53,272.52
Total population (1,000s)	322,179.61
Population growth rate (%)	0.70
Population sex ratio (female/male)	0.98
Human Capital Index score	74.84

	2006		2017	
	rank	score	rank	score
Global Gender Gap score	**23**	**0.704**	**49**	**0.718**
Economic participation and opportunity	3	0.759	19	0.776
Educational attainment	66	0.982	1	1.000
Health and survival	1	0.980	82	0.973
Political empowerment	66	0.097	96	0.124
rank out of	115		144	

COUNTRY SCORE CARD

	rank	score	avg	female	male	f/m
Economic participation and opportunity	**19**	**0.776**	**0.585**			
Labour force participation	57	0.855	0.667	66.2	77.4	0.86
Wage equality for similar work (survey)	27	0.734	0.634			0.73
Estimated earned income (PPP, US$)	56	0.648	0.509	45,287	69,901	0.65
Legislators, senior officials and managers	15	0.767	0.320	43.4	56.6	0.77
Professional and technical workers	1	1.000	0.758	57.1	42.9	1.33
Educational attainment	**1**	**1.000**	**0.953**			
Literacy rate	1	1.000	0.883	99.0	99.0	1.00
Enrolment in primary education	1	1.000	0.979	94.1	93.4	1.01
Enrolment in secondary education	1	1.000	0.971	92.0	89.0	1.03
Enrolment in tertiary education	1	1.000	0.938	99.6	72.8	1.37
Health and survival	**82**	**0.973**	**0.956**			
Sex ratio at birth	1	0.944	0.920			0.95
Healthy life expectancy	91	1.040	1.037	70.4	67.7	1.04
Political empowerment	**96**	**0.124**	**0.227**			
Women in parliament	85	0.241	0.279	19.4	80.6	0.24
Women in ministerial positions	84	0.200	0.209	16.7	83.3	0.20
Years with female head of state (last 50)	69	0.000	0.200	0.0	50.0	0.00



0.718 / 49 USA

SELECTED CONTEXTUAL DATA

Workforce Participation	female	male	value
Non-discrimination laws, hiring women			yes
Youth not in employment or education	17.4	15.6	1.11
Unemployed adults	4.8	4.9	0.97
Discouraged job seekers	37.6	62.4	0.60
Workers in informal employment	–	–	–
High-skilled share of labour force	17.5	17.8	0.98
Workers employed part-time	22.7	12.9	1.77
Contributing family workers	0.1	0.0	1.67
Own-account workers	5.1	7.4	0.69
Work, minutes per day	484.0	471.0	1.03
Proportion of unpaid work per day	50.0	31.5	1.59

Economic Leadership	female	male	value
Law mandates equal pay			no
Advancement of women to leadership roles			[2] 0.78
Boards of publicly traded companies	16.4	83.6	0.20
Firms with female (co-)owners			–
Firms with female top managers			–
Employers	–	0.0	–
R&D personnel	–	–	–

Access to Assets	female	male	value
Hold an account at a financial institution	94.8	92.4	1.03
Women's access to financial services			yes
Inheritance rights for daughters			yes
Women's access to land use, control and ownership			yes
Women's access to non-land assets use, control and ownership			yes
Mean monthly earnings (1,000s, local curr.)	0.9	1.1	0.78

Political Leadership	female	male	value
Year women received right to vote			1920
Years since any women received voting rights			97
Number of female heads of state to date			0
Election list quotas for women, national			–
Election list quotas for women, local			–
Voluntary political party quotas			–
Seats held in upper house	–	–	–

Family	female	male	value
Average length of single life	23.7	24.0	0.99
Proportion married by age 25	42.2	30.0	1.41
Mean age of women at birth of first child			30
Average number of children per woman			1.87
Women's unmet demand for family planning			8.00
Potential support ratio			4
Total dependency ratio			52
Parity of parental rights in marriage			yes
Parity of parental rights after divorce			yes

Care	female	male	value
Length of parental leave (days)			0
Length of maternity/paternity leave (days)	–	–	
Wages paid during maternity/paternity leave	–	–	
Provider of parental leave benefits			–
Provider of maternity/paternity leave benefits	–	–	
Government supports or provides childcare			yes
Government provides child allowance			yes

Education and Skills	female	male	value
Out-of-school children	5.2	5.8	0.90
Primary education attainment, adults	98.8	98.8	1.00
Primary education attainment, 25-54	–	–	–
Primary education attainment, 65+	–	–	–
Out-of-school youth	6.5	8.4	0.77
Secondary education attainment, adults	88.8	88.0	1.01
Secondary education attainment, 25-54	–	–	–
Secondary education attainment, 65+	–	–	–
Tertiary education attainment, adults	32.7	32.3	1.01
Tertiary education attainment, age 25-54	–	–	–
Tertiary education attainment, age 65+	–	–	–
PhD graduates	1.4	2.1	0.66
Individuals using the internet	74.9	74.2	1.01

Graduates by Degree Type	female	male	value
Agri., Forestry, Fisheries and Veterinary	0.8	1.1	0.73
Arts and Humanities	21.4	20.2	1.06
Business, Admin. and Law	17.3	23.4	0.74
Education	9.9	3.9	2.52
Engineering, Manuf. and Construction	2.6	13.3	0.19
Health and Welfare	22.5	7.3	3.08
Information and Comm. Technologies	1.1	6.0	0.19
Natural Sci., Mathematics and Statistics	4.9	6.6	0.74
Services	6.1	7.6	0.81
Social Sci., Journalism and Information	13.4	10.7	1.25

Health	female	male	value
Mortality, children under age 5	11.0	13.9	[1] 0.79
Mortality, non-communicable diseases	1,169.2	1,129.5	[1] 1.04
Mortality, infectious and parasitic diseases	21.5	21.8	[1] 0.99
Mortality, accidental injuries	40.7	61.2	[1] 0.66
Mortality, intentional injuries, self-harm	14.2	48.8	[1] 0.29
Mortality, childbirth			[1] –
Legislation on domestic violence			yes
Prevalence of gender violence in lifetime			36.0
Law permits abortion to preserve a woman's physical health			yes
Births attended by skilled health personnel			–
Antenatal care, at least four visits			–

[1] Age-standardized death rates per 100,000 population. [2] Data on a 0-to-1 scale (0 = worst score, 1 = best score)