# Pattern Mining

Robbe Nooyens

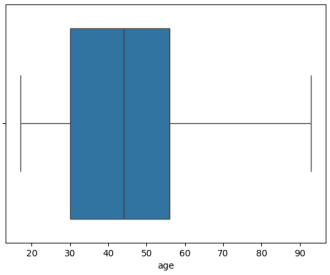
March 2024

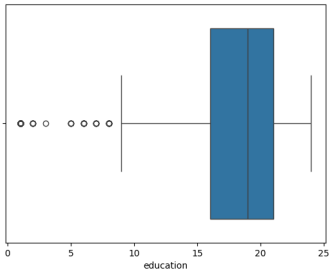
## 1 Introduction

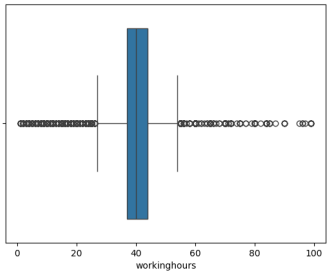
This report explores patterns in a dataset about income levels, focusing on data inspection, preparation, and analysis. By applying several data preprocessing techniques and analyzing the impact of various parameters on pattern mining outcomes, we aim to highlight key insights into the socio-economic characteristics that differentiate male and female working conditions.

## 2 Data inspection

First, we'll inspect the dataset. We can see that there are 9000 records, each having 10 columns. Two columns have missing values; around 8600 entries in 'ability to speak English' and around 7000 entries in 'gave birth this year'. We can plot the distribution of column values to properly group values:



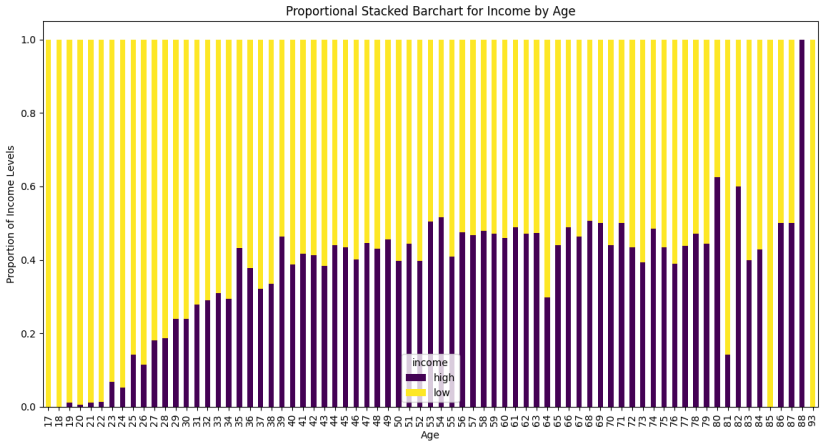




## 3 Data preparation

In order for the dataset to be consistent, the data will first be preprocessed. We'll fill the missing values in the two columns and apply some additional mapping. For the 'ability to speak English', I first filled it with 0 to represent 'Native'. Since this fills the data for 95% with the same value, a lot of nonsense and trivial rules appeared with Native as the implication. Therefore, I decided to set this to 'N/A', meaning all rules with this N/A value will be disregarded. For the 'gave birth this year', I set the default value to be 'No'. A default value for 77% of the data of 'No', also resulted in nonsense rules like 'Male' implies 'No'. Setting them to N/A removes these nonsense rules.

All fields have values now. However, there are still integer columns with a wide range. To solve this, integer values are grouped together in 'bins'. For my first setup, I decided on the bins based on the graphs from the inspection. However, labels decided on using this grouping didn't appear that much in the patterns. That's why I cut the data in less bins for the second setup based on the actual logical categories of the columns, rather than purely on the graphs and statistical data.




**Age**: When looking at the plot that represents the shares of income levels per age, we can see a rising trend per 2 columns until it stabilizes around 50%. Therefore I first grouped them per 2 years until it reached the stable state. For the second setup, I only distinguished between juniors, seniors and retired people. The amount of experience probably influences income the most. Juniors are just graduated, while seniors already have experience and are worth more in the market. A quarter of the dataset are juniors, half of the dataset is senior and another quarter is retired.

**Education**: Since most of the data is situated above a level of 15, I decided to group the first few levels. This was still too specific, which is why I just grouped them into 3 categories 'No diploma', 'No higher education' and 'Higher education'. This change gave expected rules like 'No higher education' implies 'Low income'.

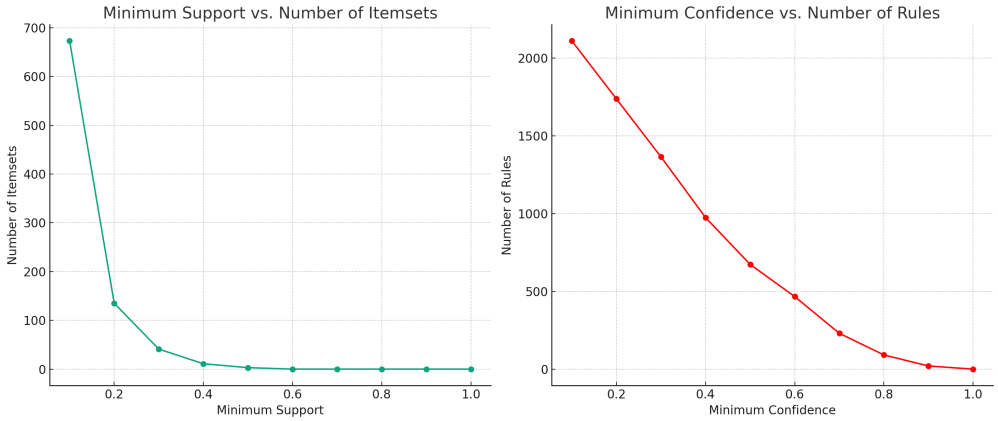
**Working hours**: Initially, I tried to be specific by using 7 categories. However, when we look at the plots, we see that the majority works 40 hours. Less or more than this is rather exceptional. That is why the second setup only has 3 categories 'Part-time', 'Full-time' and 'Overtime'.

**Language**: Four categories didn't seem much. That's why the first setup uses the 4 separate groups. However, given that only 5% is not a native English speaker, I grouped those 4 as 'Not native'. This small density means that we'd need to set a low support to see rules with 'Not native'.

**Mappings**: To make the rules as meaningful as possible, I added some optional mappings. Yes and No map to Pregnant and not pregnant and low and high map to Low Income and High Income. On top of that, I grouped all single marital statuses as 'Single', and 'Husband' and 'Wife' as 'Married' to prevent rules saying 'Husband' implies 'Male'. For the second setup, 'Not pregnant' is replaced with 'N/A' since this caused a lot of 'Not pregnant' implying 'Female' rules.

# 4 Parameter influence

Now, we'll be looking at the effects of changing the parameters for minimum support and minimum confidence. To do so, I've created several setups with one parameter fixed and the other variable. Minimum support is fixed to 0.1 and minimum confidence is fixed to 0.5. When we plot the number of results, we get the following graphs:




Both graphs show that the lower the threshold are, the more results are returned. When the minimum support threshold is close to one, very few itemsets meet the criterion. This is because a high support value means that an itemset must appear very frequently across transactions. In many datasets, only a few itemsets are that common. As the minimum support threshold decreases, more itemsets qualify as frequent, hence the number of itemsets increases. This is because lower support values include itemsets that appear less frequently. A high minimum confidence threshold results in fewer association rules being generated. High confidence requires that the implication of the rule is found in nearly all transactions where the first part appears, which is a strong condition not met by many potential rules. Lowering the confidence threshold means accepting rules where the consequent is less consistently present in transactions containing the antecedent. This allows for more rules to meet the threshold, thus increasing the number of generated rules. As confidence decreases, we see more rules, but they don't go up as quickly as itemsets do when we lower support. This is because finding rules isn't just about how often items show up together. It's also about how likely it is that one item implies another. So, even if we're less strict about confidence, it doesn't automatically mean a lot more rules pass the test, because it's also about this chance of one item leading to another, not just how many times they show up together.

# 5 Gender rules

For the sociology research, we'll try to extract rules that have Male or Female as the consequence. To do that, we'll run a setup with a low support, such that we get at least three rules per gender. There are more males than females in the dataset, which makes getting a high support for females more difficult. So in this case, we're more interested in the confidence than the support.

| Support | Confidence | Rule |
|---|---|---|
| 0.22 | 0.84 | ['Married', 'High Income'] → ['Male'] |
| 0.14 | 0.82 | ['Overtime', 'Seniors'] → ['Male'] |
| 0.27 | 0.80 | ['High Income'] → ['Male'] |
| 0.06 | 0.57 | ['Office/Administrative Support'] → ['Female', 'Low Income'] |
| 0.05 | 0.62 | ['Married', 'Part-time', 'Low Income'] → ['Female'] |
| 0.05 | 0.62 | ['Seniors', 'Part-time'] → ['Female'] |

For males, the association rules highlight a significant correlation with 'Married' and 'High Income' statuses, suggesting males in high-income brackets who are married have certain behaviors or possess specific characteristics. Rules showing a connection with 'Overtime' and 'Seniors,' suggest that older males who work overtime are more prominent than females. The higher confidence levels in these rules for males demonstrate a strong reliability in the patterns identified. The rules for females reveal a different set of associations, with lower confidence levels overall. The patterns for females are more varied, related to 'Office/Administrative Support,' 'Part-time,' and 'Low Income' statuses. The contrast between the genders is very notable. Males

are more frequently associated with high income, overtime and marital status, while females are depicted with part-time work and lower income brackets.