Data Mining Assignment 1
Frequent Pattern Mining

Maksim Karnaukh
1st year Masters at Department of Computer Science
University of Antwerp
Antwerp, Belgium
Email: maksim.karnaukh@student.uantwerpen.be

1. Introduction

We analyzed the dataset 'income_levels.xlsx'. This data consists of information of individuals from the US, aged between 17 and 93 years old. Along the age of these people the data contains information about their education level, their occupation, their marital status and so forth.

This report presents a comprehensive analysis of the dataset, including key findings and observations about the dataset features, and implications derived from the association rules extracted.

2. Data Inspection and Preparation

I tried to explore as much of the dataset as possible. The results and observations will be discussed in this section. The focus was on the five features: "ability to speak english", "gave birth this year", "age", "workinghours" and "education". I will also discuss interesting findings about other features in a shorter manner.

2.1. Handling Missing Values

Upon inspection, only two columns contained missing values in the dataset: "ability to speak english" and "gave birth this year". To address this, I implemented an appropriate imputation strategy to fill in missing values.

For the "ability to speak english" column, we work with numeric encodings of the person's ability to speak English with a scale from 1 (very well) to 4 (not at all). The field is left blank if the person is a native English speaker. To fill in the missing values here, which are the native speakers, I used a 0. This way, we maintain the ordinal nature of the scale and we ensure that the imputation is meaningful within the context of the data.

The "gave birth this year" column describes whether the person gave birth over the last 12 months (a "Yes" or "No"). Here, I filled in the missing values with a "No". First of all, men can't get pregnant so they should have a 'No' by default. For women that had a missing value there, I filled in with a 'No' as well since taking an average over the values for women wasn't a good option because we are concerned with giving birth in only the last year, and not overall (also, the amount of 'filled in' values was only 21.6% which is relatively low to draw correct conclusions from). Furthermore, and probably one of the most important reasons for the women, we can see that the missing values in the 'gave birth this year' column are only present in the 49-93 age range. Since it was proven to be a higher health risk to give birth after roughly 35 years old for women, this could be seen as a reason for why the missing values are mostly in the 38+ age bin (also, we have to keep in mind menopause which occurs roughly between the ages of 45 and 55 years). An imputation of 'No' seems to be the most logical option here. [4] [5]

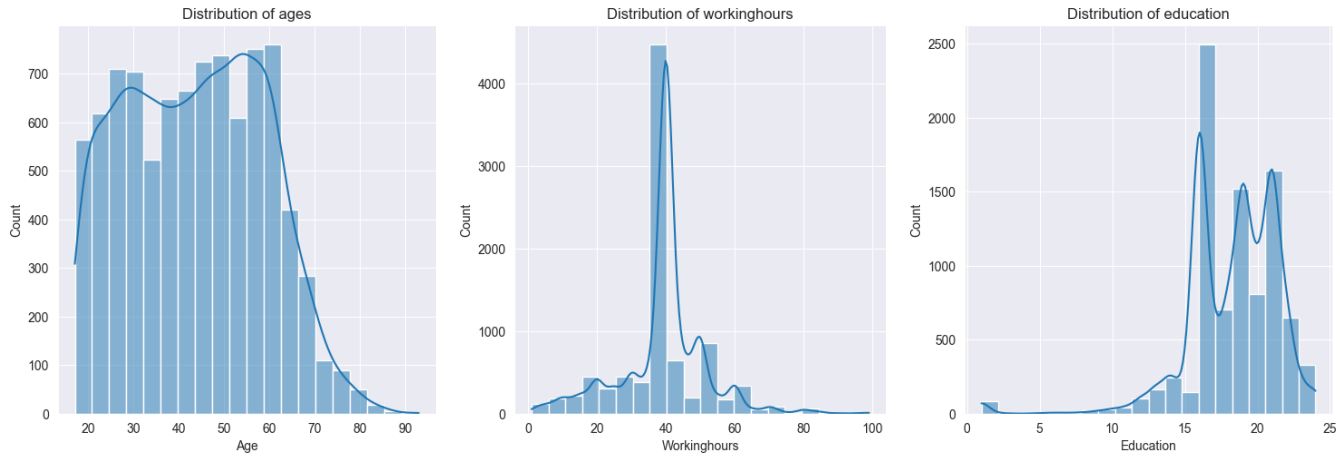
2.2. Categorizing Features

The features "age", "workinghours" and "education" have a wide range of possible values. Categorizing/binning these features will help in extracting a more concise set of patterns when we apply our frequent pattern mining algorithm.

2.2.1. Categorizing the "age" feature. Considering the "age" column, if we were to look at the distribution of the ages, we can see that the distribution has a bit of a negative skew (increasingly lower counts for 65+), with two small peaks in roughly the ranges 25-35 and 45-65.

Here, I used a bit of customized binning for the 'age' column; the ranges are: '17-28', '28-38', '38-49', '49-65', '65-93', but this was based on the quantile-based binning output. Furthermore, other reasons for this binning include the results of the average workinghours scatterplot, where we see that people in the range '17-28' and in the range '65-93' are significantly different from the other age ranges, meaning the large '65-93' group is a viable option (this will be covered in the next subsection) and the fact that the first four age ranges have a difference of roughly 10 years, which is a not a bad amount of time for a generation.

A possible downside of this exact method is the fact that the last age range ('65-93') has a significantly lower count compared to the other ranges. The good thing here is that the other ranges have similar, high value counts.



(a) (b) (c)

Figure 1: Distributions for the three features (age, workinghours and education)

2.2.2. Categorizing the ”education” feature. If we look at education, we see that most people have some form of higher education. The percentage of people that have at least a regular high school diploma is around 91%.

The categories for this column were more strictly based on the International Standard Classsification of Education (ISCED) [1]. This seemed the most logical way to approach this. The created categories are: ’Less than basic’ (0-3), ’Basic’ (4-12), ’Intermediate’ (13-20) and ’Advanced’ (21-24). The value counts are distributed very unevenly, e.g. the first two categories combined form only 3% of the total amount of values.

2.2.3. Categorizing the ”workinghours” feature. For the ”workinghours” column, it is highly unevenly distributed. Around 47% of the people work 40 hours per week, while the rest of the hours don’t reach the 10% mark.

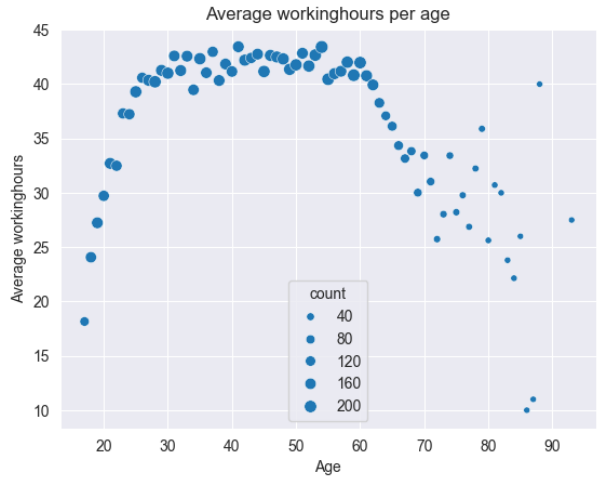


Figure 2: Average workinghours per age with their counts

To create bins for this column, I used a categorization roughly according to [3] and [6], where I created three categories: ’Part-time’ (0-30 hours), ’Full-time’ (31-40 hours) and ’Overtime’ (41-99 hours). The value counts for these ranges are a bit uneven, but not too much (’Full-time has the most values covering almost 54%, but this was expected considering the initial distribution). Such a categorization of only a few ranges is also good and easier for when we will look for association rules.

The plot in Figure 2, which we used for the ”age” binning, is also interesting here. It shows what the average working hours are per age, and how many people there are per such data point. We can see a reverse U-shape (parabola), which is quite interesting. We can observe that people on average work less when they are younger, then work more when they are in their 30s to their 60s, and then work less again when they are older. It’s noticeable from the plot that most people in our dataset are in the 40-45 working hours per week range.

2.3. Extra Insights

Other than the features discussed previously, there are some interesting things to note about the dataset which might prove to be important in the search for association rules too.

Should we look at the distribution of male to female people in the dataset, we can see that it consists of 66.7% male and 33.3% female persons. One conclusion (which will certainly prove important later) we can draw from this really quickly is that we are never going to see an association rule containing ’Female’ that has a support higher than 33.3% in the dataset.

The amount of people that are native English speakers is very high (95.5%), and the same can be said for the ’No’ value in the ”gave birth this year” column (which was 20.1% before and 98.6% after imputing the missing values with a ’No’).

After calculation, we can also note that for women, around 20% have a high income, while the percentage for men with a high income is around 41%. Also, people are mostly working in the 'private' sector (73.8%), with the second highest being 'governmental' (15.7%).

3. Search for Association Rules

Finally, we are at the second big part, which is to explore the dataset by finding frequent patterns that occur in it. I used the Apriori algorithm from [2] to extract frequent itemsets and generate association rules.

3.1. Basic Exploration with Apriori

I first ran the algorithm with different values for the "minimum support" and "minimum confidence" to see what we can notice about the number of rules we can find and the nature of these rules.

TABLE 1: Runs with different minimum support and confidence and the corresponding number of results (amount of frequent itemsets with at least two items) and association rules.

run	min. sup.	min. conf.	# freq. itemsets	# rules
1	0.2	0.95	154	248
2	0.6	0.95	13	20
3	0.8	0.9	1	2
4	0.8	0.2	1	2
5	0.6	0.5	13	42
6	0.2	0.2	207	1942

From the above table, we can notice that the lower our minimum support, the more rules we find. The same applies to the minimum confidence. It is however so that an increase in minimum support has a bigger impact on the number of rules than the minimum confidence as we can see in figure 5 (the exact impact of changing these thresholds for a certain dataset can vary depending on the dataset's characteristics).

In general, lowering the minimum support will (typically) result in a larger number of frequent itemsets and association rules. This is because more itemsets will meet the lower support threshold, leading to more potential association rules. Reducing the minimum confidence will lead to more association rules being generated. Lower confidence thresholds allow weaker associations to be considered significant, resulting in a larger number of rules.

As for the nature of the generated rules, those generated with lower support and confidence thresholds mostly include less meaningful associations. These rules may be less reliable. Rules generated with higher support and confidence thresholds represent strong (and meaningful) associations. These rules tend to be more reliable for decision-making.

We see that rules with a confidence of 1.00 are quite straightforward, e.g. if the person is a "Male", then he/she has not "given birth this year" or if the person is a "Husband", then he is a "Male". In fact, every rule with confidence 1.00 has either one of these terms or something similar. We also see quite logical rules with high confidence,

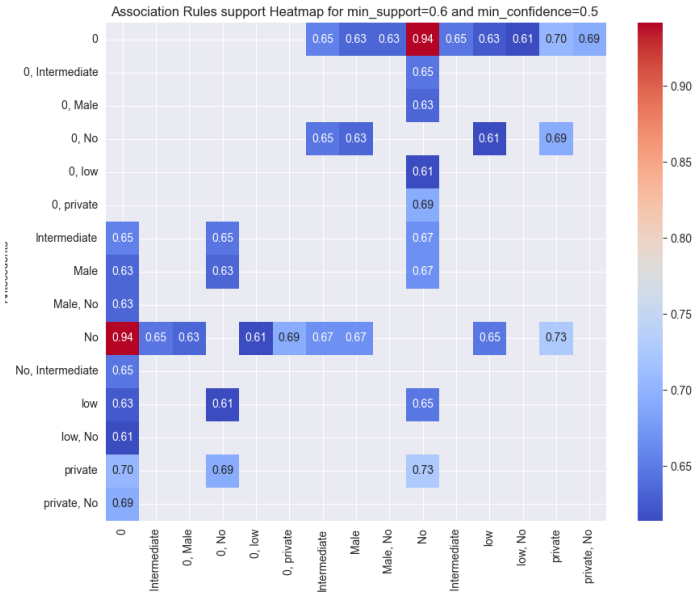


Figure 3: Support heatmap for the association rules generated by the apriori algorithm with antecedents on the y-axis.

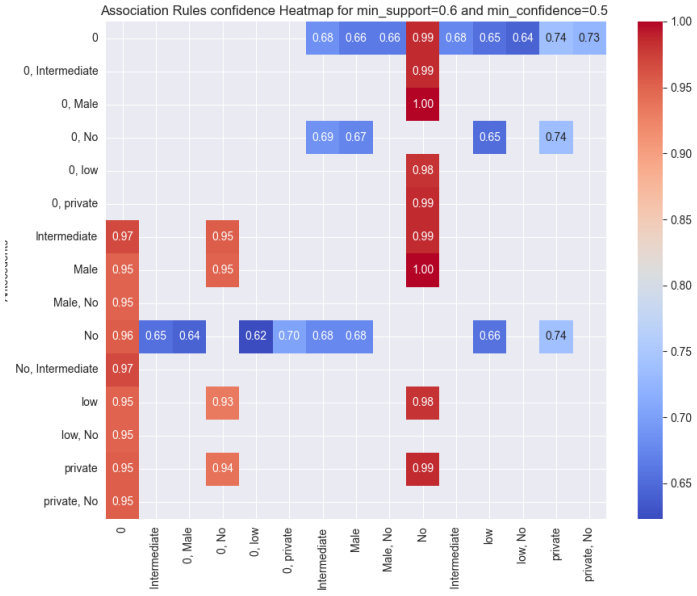


Figure 4: Confidence heatmap for the association rules generated by the apriori algorithm with antecedents on the y-axis.

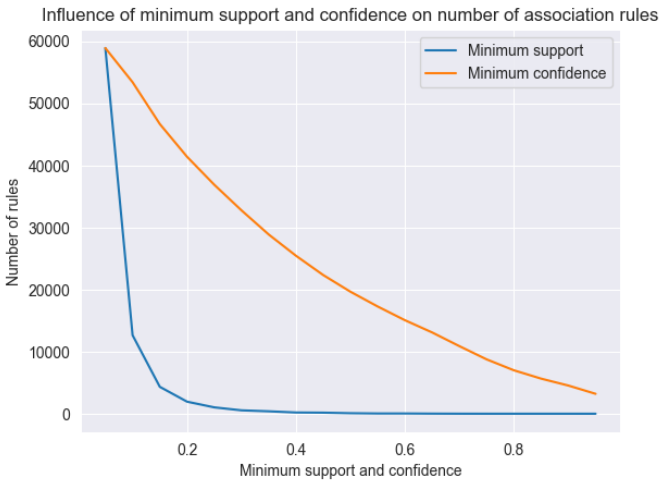


Figure 5: Plot to show influence of varying minimum support with fixed minimum confidence of 0.05 and vice versa on the total number of resulting association rules.

e.g. if the person was "Never married", he/she didn't "give birth this year".

The highest support for a rule containing "Male" is 0.66. For "Female", this is 0.32. This is understandable since there are 66% "Males" and 33% "Females" in the dataset. For potential future section references, I will call this the male bias in the dataset.

3.2. Male and Female Rules

Next, we are going to try to study the population differences in Men and Women. We extract rules that have "sex = Male" or "sex = Female" as their consequence. We will try to find rules with high support, high confidence or both and describe some patterns that we have found (for each sex at least 3 distinct ones).

First of all, it is important to note that since we have the male bias in our dataset, no rule with "Male" as a consequence can possibly have a support higher than 0.67 and no rule with "Female" as a consequence can possibly have a support higher than 0.33. In fact, there are no rules at all with a support higher than 0.4. I tried splitting the dataset into an only-male dataset and only-female dataset, but the reason I didn't further use this is because then the confidence factor loses most of its meaning.

In Table 2, I put for each sex at least three distinct rules from which we can describe patterns. When I was looking at a large subset of rules with "sex = Male" or "sex = Female" in their consequence, I noticed that on average the support for the male rules was higher and the same applies to the confidence (the formula for confidence uses the support values). This is likely because of the male bias. Partly because of this, the confidence is usually a bit more informative. It is also the case that some columns have a larger amount of different values which don't have very high value counts, meaning the support there will naturally be quite low.

I would like to quickly discuss the lift of the rules. In the assignment it wasn't really mentioned. Since determining what a good support and confidence value for a rule is and drawing conclusions from these can be ambiguous and hard (a lot of people have different opinions on this), people have come up with more ways to describe the properties of a rule. One of them is the lift, which is the ratio between the confidence and support expressed as:
$Lift(A \rightarrow B) = (Confidence(B \rightarrow A))/(Support(A))$.
This can be seen as the (prediction) strength of the rule. The value of lift is that it considers both the confidence of the rule and the overall data set. Lift of greater than 1 means that items are more dependent (while a lift of 1 means there is no association between the items).

TABLE 2: Male/Female patterns

Nr.	Rule	Support	Confidence	Lift
1	high → Male	0.274	0.801	1.201
2	Overtime → Male	0.220	0.813	1.219
3	Construction/Extraction → Male	0.067	0.977	1.46590
4	Management/Business → Male	0.086	0.730	1.095
5	Advanced → Male	0.185	0.637	0.955
6	Office/Admin. Support → Female	0.064	0.667	2.000
7	low → Female	0.265	0.403	1.209
8	Full-time → Female	0.191	0.355	1.065
9	Advanced → Female	0.106	0.363	1.090

For the table I used rules with lift values greater than 1. This doesn't apply to rule 5 which is why I won't use it to describe a pattern (this serves as an example to not blindly pick something with decent support and confidence).

As we can see, if a person earns a high salary, he's likely to be a male. If the person works overtime, he's also likely a male. Interestingly also, and this is not visible in the table, is that if the person earns a high income *and* works overtime, the probability of the person to be a male is higher (0.863) with a support of 0.125 and lift of 1.296. If someone works in construction/extraction, we can almost for sure say that he's a male. For the management/business sector the confidence is slightly lower, but still relatively high. Based on (probably wrong sexist) stereotypes, I did expect all the above in a way. Keep in mind that although for example the supports for rule nr. 1 and 2 appear low, they are relatively high (relative to the percentage of males in the dataset, the percentage of high (vs. low) income and the percentage of the "Overtime" category).

If someone works in Office/Administrative Support, it is more likely to be a women (very high lift value by the way). For a low income and Full-time work, the support is quite high, but the confidence is rather low. This is normal though. I would expect to actually maybe see more women in Advanced education than men. We see that we have a support of 0.106 there (for male it's 0.185 with 0.637 confidence). If we however take the relative support we get a support of around 0.32, while for male the relative support

is 0.28. The thing is that since there is a male bias, the confidence values for the female rules suffer because of this bias.

4. Notes

The code and plots/pictures for this project can be found at https://github.com/MaksimKarnaukh/DataMining.

References

[1] ILOSTAT. International standard classsification of education (isced), 2023. https://ilostat.ilo.org/resources/concepts-and-definitions/classification-education/.

[2] Yu Mochizuki. Apyori 1.1.2, 2019. https://pypi.org/project/apyori/.

[3] U.S. Bureau of Labor Statistics. Labor force, employment, and unemployment concepts, 2023. ://www.bls.gov/cps/definitions.htmfullparttime.

[4] The American College of Obstetricians and Gynecologists. Having a baby after age 35: How aging affects fertility and pregnancy, 2023. https://www.acog.org/womens-health/faqs/having-a-baby-after-age-35-how-aging-affects-fertility-and-pregnancy#:~:text=The%20risks%20of%20miscarriage%20and,chance%20of%20a%20multiple%20pregnancy.

[5] World Health Organization. Menopause, 2022. https://www.who.int/news-room/fact-sheets/detail/menopause#:~:text=Most%20women%20experience%20menopause%20between,changes%20in%20the%20menstrual%20cycle.

[6] Glassdoor Team. Exactly How Many Hours Is Considered Part-Time?, 2021. https://www.glassdoor.com/blog/guide/how-many-hours-is-part-time/.