Association Mining on Wages Dataset:
What Drives Value?

Theodore Fitch

Department of Data Analytics, University of Maryland Global Campus

DATA 630: Machine Learning

Dr. Ami Gates

June 8th, Summer 2021

Introduction:

The purpose of this analysis is to find patterns within a dataset pertaining to wages using the Apriori method. This method entails finding correlations within data and listing them based on how good they are. The correlations are rules of probability where 1 item is predicted given a precondition (i.e. given X has already occurred, what is the probability of Y). Each rule is measured by 3 values: 1. Support – the percentage of rows supporting such a rule with both X and Y (X and Y are also referred to as Left Hand Side [LHS] and Right Hand Side [RHS] 2. Confidence – the probably of Y given X 3. Lift – Support/P(X) (or Support divided by the probability of X), which entails that a value > 1 is positively correlated and a value < 1 is negatively correlated (Han et al., 2011).

Wage rates affect everyone. They feel personal because they're a reflection of how much value a person brings to a company, and they determine our standard of living. In particular, wages are a current, hot topic where there are calls to increase the federal minimum wages to $15 per hour (Reich, 2019) and discussion around the gender pay gap (Blau & Khan, 2020). From a more meta perspective, wage rates also touch on the topics of supply and demand, federal interventionism, and the economic health of a nation. It's an incredible thing of the current time where the vast majority of people live above the absolute poverty line (Allen, 2017). In fact, absolute poverty may completely disappear by the 2040s with the rapid technological and financial gains made in recent years (Allen, 2017). That said, to mandate a federal minimum wage has major implications: a company may outsource many roles to other countries where those minimum wage rates don't exist; some states or people may think it's a governmental overreach to mandate wage rates and push back locally; and it could greatly increase the quality of life for people who would then earn more. But predicting how mandating wage rates would affect society is first predicated on knowing what wage rate disparities exist and why they exist.

Using the aforementioned apriori method can help find useful associations related to these topics. This knowledge has impact in quite a few areas including economic policy at the government level and worker pay at the corporate level. The Apriori method is excellent at determining correlations. It cannot state definitive causes, but it can say how certain variables may be related and how strongly. Other tools can later be used to determine causal relationships (like regression). It's expected this method should highlight general trends in the wage rate dataset, finding unique correlations regarding wages and other factors like race, sex, years of experience, and union status. In particular, it would be interesting to find trends regarding high wages or low wages so that those variables could be explored further to determine if they are causal or not.

Analysis and Model Demonstration:

Data Information:

The dataset used was created by pulling 534 random rows from the Current Population Survey (CPS) in 1985. The CPS is taken between census years in order to garner more data (Determinants, n.d.). Originally, this dataset was used to determine if there were any correlations between the other variables and wages, and to determine if there is a gender wage gap.

Exploratory Data Analysis:

The "Wages" dataset contains 534 rows of data amongst 11 variables (Figure 1). There are 4 binary variables: South, Sex, Union, and Marital Status (a 1 in these columns respectively denote that a person: lives in the south; is female; is a union member; is married). There are 3 factor variables: Race (where 1 = Other, 2 = Hispanic, and 3 = White), Occupation (where 1 =

Management, 2 = Sales, 3 = Clerical, 4 = Service, 5 = Professional, 6 = Other), and Sector (where 0 = Other, 1 = Manufacturing, 2 = Construction). All of the aforementioned variables are qualitative in nature. Lastly, there are 4 ratio variables (which are inherently quantitative): education (in years), experience (in years), wage (in dollars per hour), and age (in years). All variables were explored, and a few variables were graphed to understand the distribution of data. There are a similar number of both sexes represented in the data (Figure 2) but there are almost double the number of married individuals than unmarried individuals (Figure 3). Wages shows a right skewed distribution with one outlier sitting at 44.5 dollars per hour (Figure 4).

Preprocessing:

Several steps were taken to pre-process the data. First, it was determined there were no null values in the dataset. Next, the data was examined for a unique/identifying variable which it was determined to not contain. The apriori method requires that any unique key or identifying variable is removed. Since there were 4 binary variables and 3 factor variables, these were all transformed into factor variables (so that they wouldn’t be recognized as numbers but as categories). The 4 ratio variables were then discretized to smooth out the data. Age, years of experience, and wage were all discretized using the interval method meaning the bins used were equidistant from one another and all the same size. However, education used the fixed method to account for different education levels. So, bin 1 is < 12 years (representing not completing high school), bin 2 is 12-14 years representing at least completing high school), bin 3 is 14-16 representing at least a few years of college, and bin 4 is > 16 representing at least an undergraduate degree. It’s important to note here that all bins increase in size and in bin number such that bin 1 is the lowest bin and contains the lowest values for age, years of experience, wage, and education.

Models and Methods:

The preprocessed dataset was first explored using the apriori method without any parameter specifications (this will be referred to as model 1)(Figure 5). This yielded a ruleset of 853 rules, where the top two rules showed that >80% of the values were not unionized and > 80% were white. The rule with the highest confidence (>98%) predicts that those who are in the youngest age bin will also belong to the lowest experience bin. This ruleset was created to find interesting general trends.

Next, the apriori method was rerun with minimum support of 0.4 and minimum confidence of 0.7 (referred to as model 2)(Figure 6). This yielded 46 rules and showed the majority (>70%) of individuals were not in the south and >76% of individuals worked in the sector “other”. Another method was run with a minimum support of 0.1, minimum confidence of 0.7, and a requirement that all RHS values must pertain to wages (referred to as model 3)(Figure 7). It yielded 63 rules and all rules pertained to bin 1 of wages ($1-8.25)(Figure 7). When exploring why only rules pertaining to bin 1 were found, it was discovered that most rules pertaining to bins 2-6 of wages had a confidence < 0.5. This is primarily due to how many values were in each bin. Bin 1 had 285, whereas the subsequent bins had 196, 41, 11, 0, and 1 value. Because the majority of values were on the low end of this distribution, general trends regarding high wage earners likely do not have a large enough sample size to be significant. Thus, the majority of efforts were focused into understanding trends in the first bins.

The rules were then sorted by lift and pruned based upon redundancy (Figure 8, 9). This pruned list shortened the ruleset from 63 rules to 35 rules (referred to as model 4)(Figure 12). These rules all have a lift of > 1.30 and a confidence > 0.70 (Figure 10)(these rules also had to pertain to wages in the RHS [Figure 7]). A majority of the rules on the LHS pertain to: the sector being other; age bin 1 (the youngest); experience bin 1 (the most inexperienced); and being non-union (Figure 11). Some less common but still notable rules which are correlated with being in the lowest wage bracket are: having a job in service; being female; being white; being in the south; being unmarried; and being in education bin 2 (12-14 years of education – equivalent to having a high school degree and maybe some college experience)(Figure 11, Figure 12).

Results and Model Evaluation:

Four models were presented: 1. A model with no requirements and no pruning 2. A model requiring a minimum of 0.4 support and 0.7 confidence 3. A model requiring a minimum of 0.4 support, minimum of 0.7 confidence, and wages in the RHS 4. A model requiring a minimum of 0.4 support, minimum of 0.7 confidence, wages in the RHS, the list was sorted in descending order by lift, and pruned so no redundant rules were included. Each model was iterative and built up from the previous model. Each model brought understanding and value to the table, but model 4 helped to answer the pertinent questions the most.

Model 1 showed that the majority of the dataset values were non-union and white (Figure 5). Other trends showed there was a similar ratio of males to females, but nearly double married individuals versus unmarried individuals (Figure 2, Figure 3). These trends help establish a baseline when examining later trends. If 80% of values are non-union, this needed to be considered when a rule is stated that non-union individuals tend to be in the lowest wage bracket at a rate of 80%. This would entail that the rate of non-union is congruent in that subset with the larger dataset. In this case, the major rules derived from model 1 will be discussed within this framework.

As previously mentioned, age bin 1 (the youngest) and experience bin 1 (the least experienced) were associated with wage bin 1 (the lowest wage). These seem to make sense as the younger a person is, the less experience they have. The less experience someone has, the less they will be compensated. Next, non-union was correlated with wage bin 1; however, non-union also naturally occurs at a rate of 80% in the whole dataset. Meaning, this rule should only be examined with other LHS rules since it occurs so high (and doesn't appear as a single LHS rule at a rate above 80%). Next, service jobs were correlated with wage bin 1. This is congruent with common sense since service jobs tend to require little to no higher education or technical training. On the other hand, fields like management, professional, sales, or clerical (some of the other listed categories) have high earning potential and/or typically require higher education or technical skills. Being female was also correlated with wage bin 1. This is unique since there are slightly more males in this dataset than females meaning this rule isn't due to unequal sampling of sexes. This would be a fascinating trend to look into for future work to understand why there is a high correlation between these two. This could be due to unfair treatment or it could be due to underlying variables (for example if there was also a high correlation between female and working in the service industry which inherently pays less than other fields). It was observed that being female and being non-union were highly correlated (89%)(Figure 6).

Being white was also correlated with wage bin 1; however, it's critical to point out this dataset is over 80% white and thus it should be treated like the rule regarding unions. If it's paired with other rules and/or exists where the confidence rate is significantly different than that of the naturally occurring rate, then it can be considered significant. More analyses would need to be performed to find this out. Next, being from the south was correlated with wage bin 1 which is also unique because this trait exists at a rate of 29% in this dataset. This makes the trend significant – though it's important to note that this rule only appeared when paired with other LHS rules (meaning it's not a very strong correlation). Being unmarried (which occurred in the dataset at a rate of 35%) was also correlated with wage bin 1; Yet, like being from the south, this rule only appeared with other rules meaning the correlation alone was not very strong. Lastly, model 4 correlated education bin 2 (12-14 years of education) with wage bin 1 (Figure 11). This naturally makes sense because those with high school degrees and/or some college with no degree are likely to make less than those with undergraduate or graduate degrees.

Thus, the top rules from model 4 all had 3 variables commonly correlated with lower wages: low education, low experience, and low age. These all tie into one another and explain why someone would have lower wages. However, there were several correlations found that are not simply explained and should be looked into further (see Limitations and Improvements section below).

Conclusion:

It has thus been shown from the wages dataset that there are several preconditions strongly correlated with lower wages: low age, low experience, and low education. Four apriori models were made and were iterations of one another starting with no specifications and ending with a non-redundant, pruned ruleset pertaining only to wages. While all 4 models brought value and helped to extract trends from the data, model 4 was the most helpful in finding trends, and had the most accurate rules (support > 0.1, confidence >0.7, and lift > 1.3). Looping back to the opening discussion, some disparities were found in this data. It was found that females and non-union members tend to make less than their counterparts. This is a correlation, and it’s not to say these are causal reasons. On the other hand, it’s expected that those who have little experience and little education would have low wages. Further studies could be performed.

Limitations and Improvements:

There were three factors correlated with low wages (wage bin 1) which would be compelling to study more: being female, being white, and being non-union. Future work around these topics would be first to find a larger dataset since the racial categories and union categories were uneven. Having similarly sized representations of categories is necessary to properly assess data and find trends. Next, it would be interesting to create apriori models where each one of these categories is in the RHS or LHS requirement. This would show if these were correlated with wage bin 1 again, or if there are associated with other trends which do drive lower wages. For example, low age, low experience, work in the service sector were all determined to be strongly, independently associated with lower wages. If one of those were also tightly correlated with one of these others (like being non-union), then it may be assumed that having low experience was the driver and being non-union was simply coincidence.

It should also be mentioned that one limitation in the original dataset was that a few categories needed to have more details. For example, race only contained White, Hispanic, and Other. Sector only contained Manufacturing, Construction, and Other. Both of these variables should have the "Other" category expanded by several factors so that more trending can be performed. Time was also a constraint in performing this analysis. Having more time would have allowed for performing an in-depth, multivariate analysis (in particular at items like sex, race, and union status). Lastly, these models focused solely on low wage rates because there were only a few values in the high wage rate bins. More data would need to be gathered to find rules regarding high wages in order to make significant claims.

References:

Allen, R. C. (2017). Absolute poverty: When necessity displaces desire. American Economic Review, 107(12), 3690-3721.

Blau, F. D., & Kahn, L. M. (2020). The gender pay gap: Have women gone as far as they can? (pp. 345-362). Routledge.

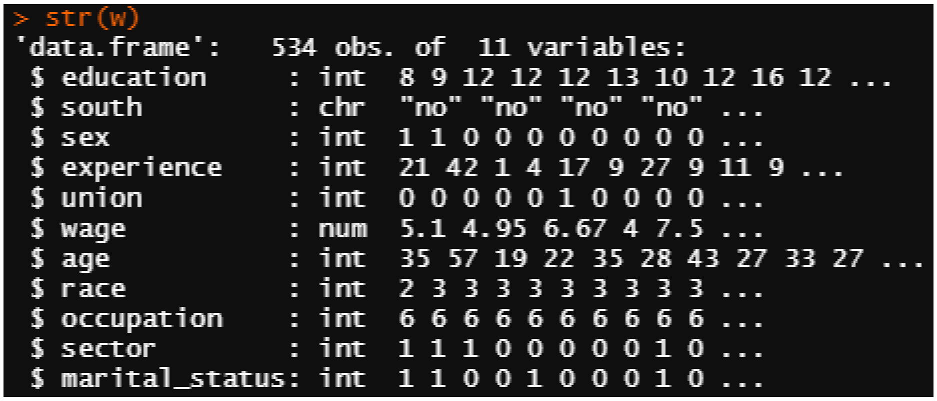
Determinants of Wages from the 1985 Current Population Survey. (n.d.). StatLib. Carnegie Mellon University. Retrieved June 6th, 2021 from: http://lib.stat.cmu.edu/datasets/CPS_85_Wages

Han, J., Kamber, M., and Pei, J. (2011). Data Mining: Concepts and Techniques, Third Edition. Elsevier. Retrieved June 6th, 2021 from:

http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

Reich, M. (2019). Likely Effects of a $15 Federal Minimum Wage by 2024. Policy Report, Center on Wage and Employment Dynamics, Institute for Research on Labor and Employment (Berkeley: University of California.

Appendix:



```
> str(w)
'data.frame':   534 obs. of  11 variables:
 $ education     : int  8 9 12 12 12 13 10 12 16 12 ...
 $ south         : chr  "no" "no" "no" "no" ...
 $ sex           : int  1 1 0 0 0 0 0 0 0 0 ...
 $ experience    : int  21 42 1 4 17 9 27 9 11 9 ...
 $ union         : int  0 0 0 0 0 1 0 0 0 0 ...
 $ wage          : num  5.1 4.95 6.67 4 7.5 ...
 $ age           : int  35 57 19 22 35 28 43 27 33 27 ...
 $ race          : int  2 3 3 3 3 3 3 3 3 3 ...
 $ occupation    : int  6 6 6 6 6 6 6 6 6 6 ...
 $ sector        : int  1 1 1 0 0 0 0 0 1 0 ...
 $ marital_status: int  1 1 0 0 1 0 0 0 1 0 ...
```

Figure 1. Structure of the wages dataset shows there are 11 variables and 534 rows of data.

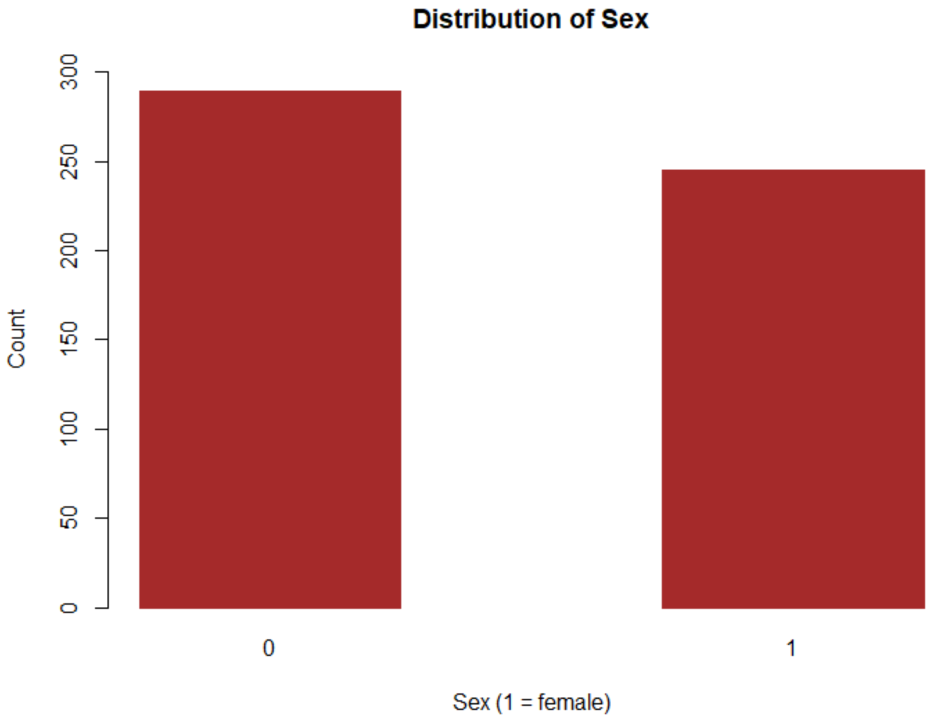


Figure 2. Distribution of sex shows there are 289 males and 245 females.

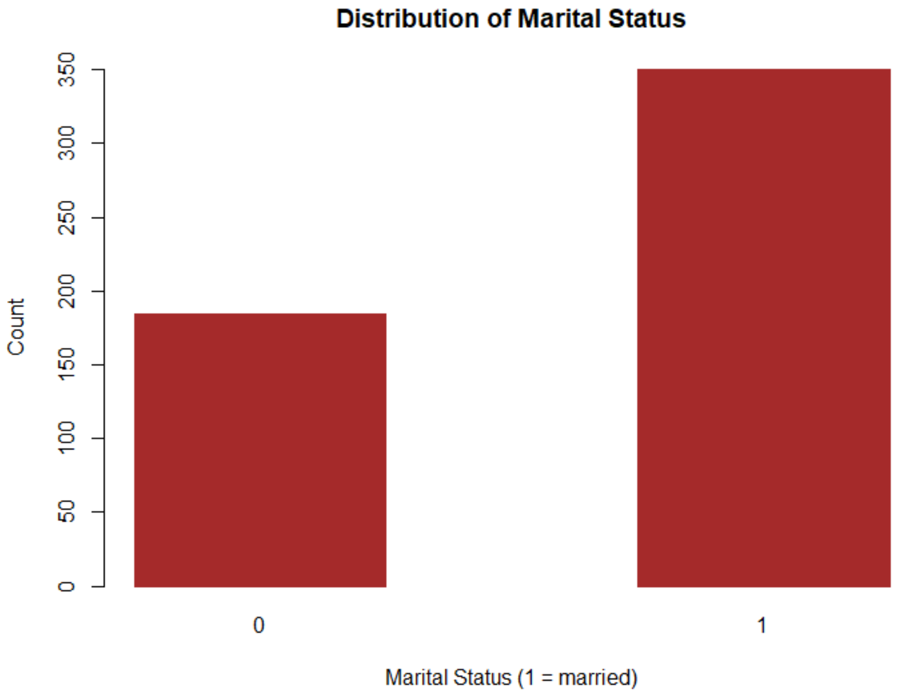


Figure 3. Distribution of marital status shows there are 184 unmarried individuals and 350 married individuals.

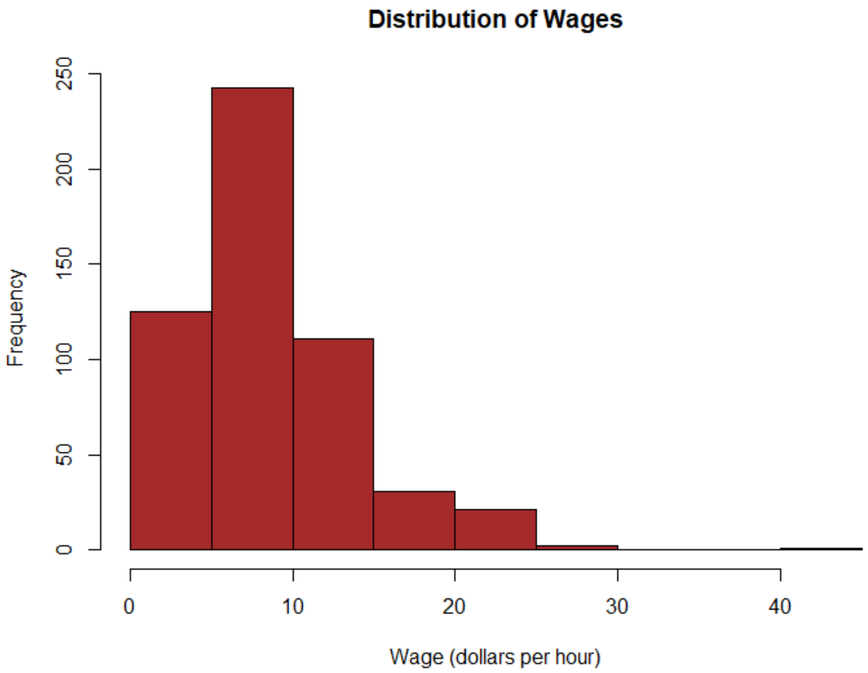


Figure 4. Distribution of wages is right skewed with 1 outlier (44.5).

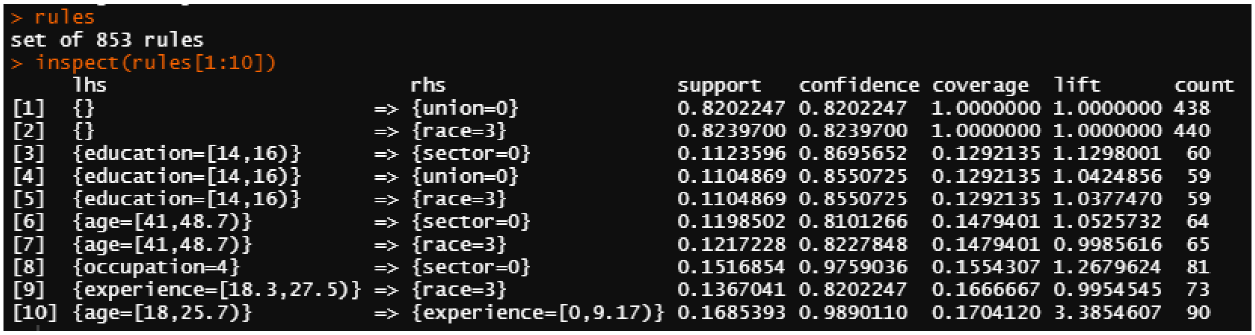

```
> rules
set of 853 rules
> inspect(rules[1:10])
     lhs                             rhs                         support   confidence coverage  lift      count
[1]  {}                           => {union=0}                   0.8202247 0.8202247  1.0000000 1.0000000 438
[2]  {}                           => {race=3}                    0.8239700 0.8239700  1.0000000 1.0000000 440
[3]  {education=[14,16)}          => {sector=0}                  0.1123596 0.8695652  0.1292135 1.1298001  60
[4]  {education=[14,16)}          => {union=0}                   0.1104869 0.8550725  0.1292135 1.0424856  59
[5]  {education=[14,16)}          => {race=3}                    0.1104869 0.8550725  0.1292135 1.0377470  59
[6]  {age=[41,48.7)}              => {sector=0}                  0.1198502 0.8101266  0.1479401 1.0525732  64
[7]  {age=[41,48.7)}              => {race=3}                    0.1217228 0.8227848  0.1479401 0.9985616  65
[8]  {occupation=4}               => {sector=0}                  0.1516854 0.9759036  0.1554307 1.2679624  81
[9]  {experience=[18.3,27.5)}     => {race=3}                    0.1367041 0.8202247  0.1666667 0.9954545  73
[10] {age=[18,25.7)}              => {experience=[0,9.17)} 0.1685393 0.9890110  0.1704120 3.3854607  90
```

Figure 5. The first unedited, unpruned set of rules yields 853 individual rules (model 1).

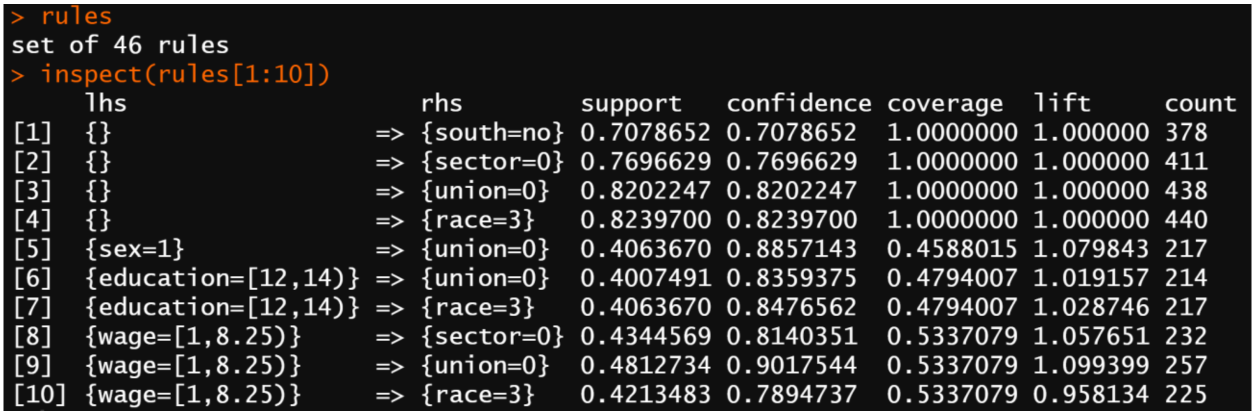
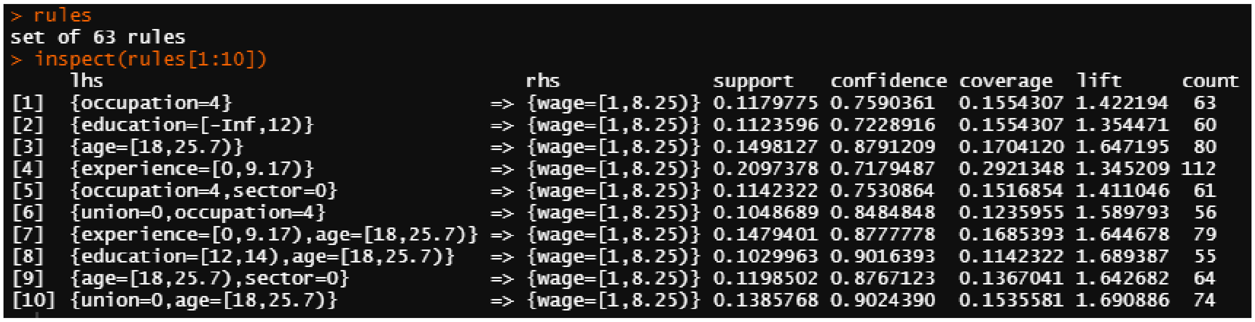
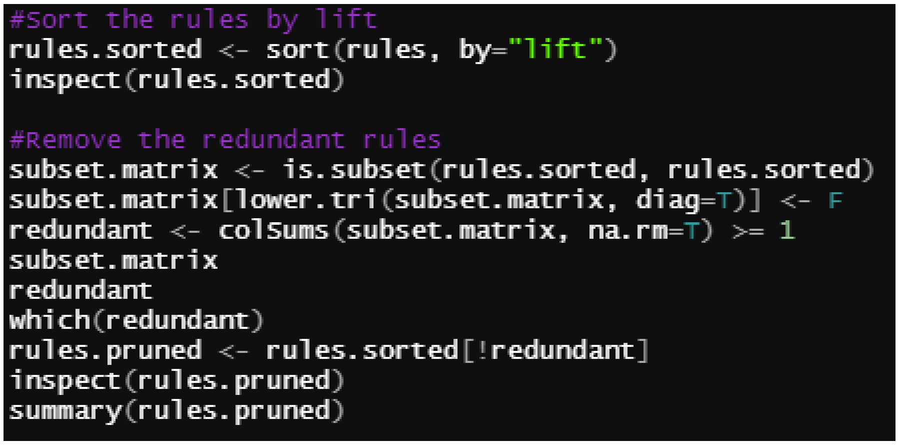

```
> rules
set of 46 rules
> inspect(rules[1:10])
     lhs                   rhs          support   confidence coverage  lift     count
[1]  {}                 => {south=no}  0.7078652 0.7078652  1.0000000 1.000000 378
[2]  {}                 => {sector=0}  0.7696629 0.7696629  1.0000000 1.000000 411
[3]  {}                 => {union=0}   0.8202247 0.8202247  1.0000000 1.000000 438
[4]  {}                 => {race=3}    0.8239700 0.8239700  1.0000000 1.000000 440
[5]  {sex=1}            => {union=0}   0.4063670 0.8857143  0.4588015 1.079843 217
[6]  {education=[12,14)} => {union=0}  0.4007491 0.8359375  0.4794007 1.019157 214
[7]  {education=[12,14)} => {race=3}   0.4063670 0.8476562  0.4794007 1.028746 217
[8]  {wage=[1,8.25)}    => {sector=0}  0.4344569 0.8140351  0.5337079 1.057651 232
[9]  {wage=[1,8.25)}    => {union=0}   0.4812734 0.9017544  0.5337079 1.099399 257
[10] {wage=[1,8.25)}    => {race=3}    0.4213483 0.7894737  0.5337079 0.958134 225
```

Figure 6. Generating a ruleset requiring a minimum support of 0.4 and confidence of 0.7 yields 46 rules (model 2).



```
> rules
set of 63 rules
> inspect(rules[1:10])
     lhs                                   rhs               support   confidence coverage  lift     count
[1]  {occupation=4}                     => {wage=[1,8.25)} 0.1179775 0.7590361  0.1554307 1.422194  63
[2]  {education=[-Inf,12)}              => {wage=[1,8.25)} 0.1123596 0.7228916  0.1554307 1.354471  60
[3]  {age=[18,25.7)}                    => {wage=[1,8.25)} 0.1498127 0.8791209  0.1704120 1.647195  80
[4]  {experience=[0,9.17)}              => {wage=[1,8.25)} 0.2097378 0.7179487  0.2921348 1.345209 112
[5]  {occupation=4,sector=0}            => {wage=[1,8.25)} 0.1142322 0.7530864  0.1516854 1.411046  61
[6]  {union=0,occupation=4}             => {wage=[1,8.25)} 0.1048689 0.8484848  0.1235955 1.589793  56
[7]  {experience=[0,9.17),age=[18,25.7)} => {wage=[1,8.25)} 0.1479401 0.8777778  0.1685393 1.644678  79
[8]  {education=[12,14),age=[18,25.7)}  => {wage=[1,8.25)} 0.1029963 0.9016393  0.1142322 1.689387  55
[9]  {age=[18,25.7),sector=0}           => {wage=[1,8.25)} 0.1198502 0.8767123  0.1367041 1.642682  64
[10] {union=0,age=[18,25.7)}            => {wage=[1,8.25)} 0.1385768 0.9024390  0.1535581 1.690886  74
```

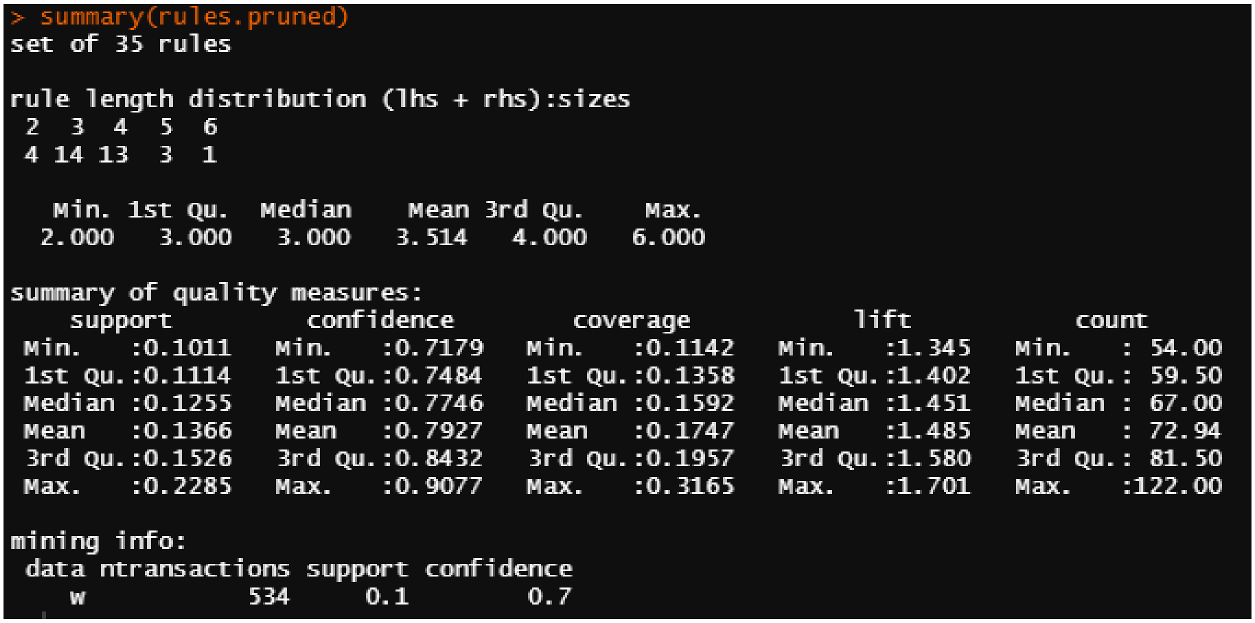
Figure 7. Generating a ruleset requiring a minimum support of > 0.1, a confidence of > 0.7, and the RHS must pertain to wages yields 63 rules (model 3).



```
#Sort the rules by lift
rules.sorted <- sort(rules, by="lift")
inspect(rules.sorted)

#Remove the redundant rules
subset.matrix <- is.subset(rules.sorted, rules.sorted)
subset.matrix[lower.tri(subset.matrix, diag=T)] <- F
redundant <- colSums(subset.matrix, na.rm=T) >= 1
subset.matrix
redundant
which(redundant)
rules.pruned <- rules.sorted[!redundant]
inspect(rules.pruned)
summary(rules.pruned)
```

Figure 8. Code used to sort ruleset by lift and prune ruleset based upon redundancy.



```
> summary(rules.pruned)
set of 35 rules

rule length distribution (lhs + rhs):sizes
 2  3  4  5  6 
 4 14 13  3  1 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.000   3.000   3.000   3.514   4.000   6.000 

summary of quality measures:
    support          confidence         coverage            lift           count       
 Min.   :0.1011   Min.   :0.7179   Min.   :0.1142   Min.   :1.345   Min.   : 54.00  
 1st Qu.:0.1114   1st Qu.:0.7484   1st Qu.:0.1358   1st Qu.:1.402   1st Qu.: 59.50  
 Median :0.1255   Median :0.7746   Median :0.1592   Median :1.451   Median : 67.00  
 Mean   :0.1366   Mean   :0.7927   Mean   :0.1747   Mean   :1.485   Mean   : 72.94  
 3rd Qu.:0.1526   3rd Qu.:0.8432   3rd Qu.:0.1957   3rd Qu.:1.580   3rd Qu.: 81.50  
 Max.   :0.2285   Max.   :0.9077   Max.   :0.3165   Max.   :1.701   Max.   :122.00  

mining info:
 data ntransactions support confidence
    w           534     0.1        0.7
```

Figure 9. Output of the summary of pruned ruleset (model 4).

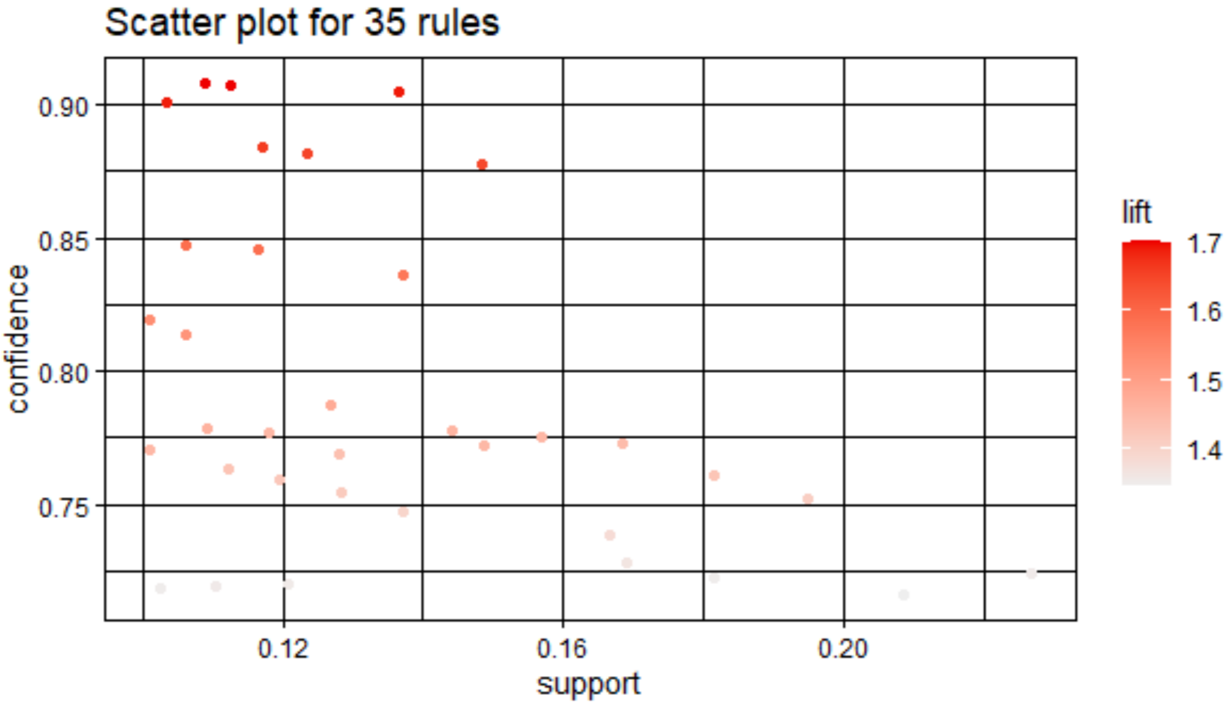


Figure 10. The pruned dataset shows a majority of rules have moderate to high confidence with a support ranging 0.01-0.24 and all lift is > 1.3 (all RHS pertain to wages)(model 4).

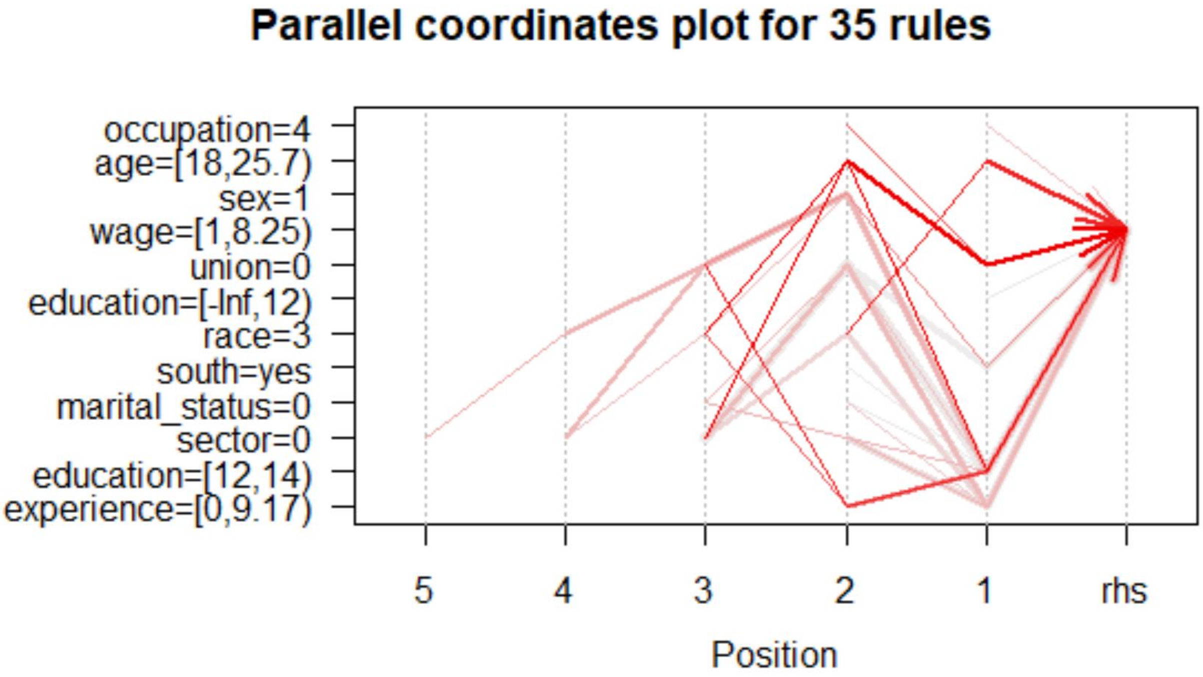
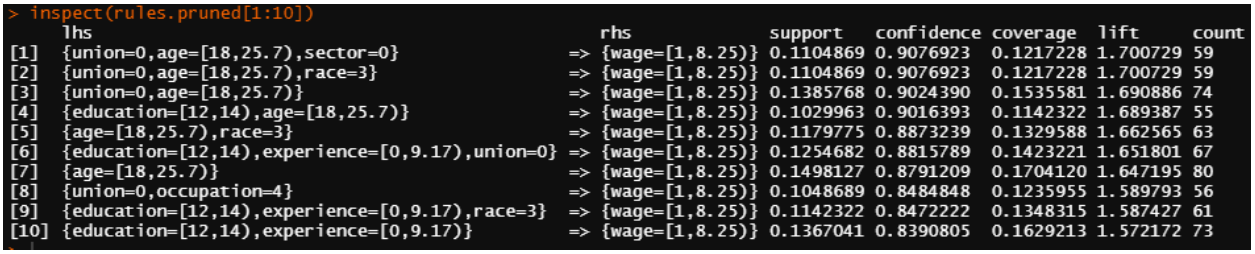


Figure 11. A parallel coordinates plot of the pruned dataset shows where a majority of rules originate from (and all RHS pertain to wages)(model 4).



```
> inspect(rules.pruned[1:10])
     lhs                                                   rhs               support   confidence coverage  lift     count
[1]  {union=0,age=[18,25.7),sector=0}                   => {wage=[1,8.25)} 0.1104869 0.9076923  0.1217228 1.700729 59
[2]  {union=0,age=[18,25.7),race=3}                     => {wage=[1,8.25)} 0.1104869 0.9076923  0.1217228 1.700729 59
[3]  {union=0,age=[18,25.7)}                            => {wage=[1,8.25)} 0.1385768 0.9024390  0.1535581 1.690886 74
[4]  {education=[12,14),age=[18,25.7)}                  => {wage=[1,8.25)} 0.1029963 0.9016393  0.1142322 1.689387 55
[5]  {age=[18,25.7),race=3}                             => {wage=[1,8.25)} 0.1179775 0.8873239  0.1329588 1.662565 63
[6]  {education=[12,14),experience=[0,9.17),union=0} => {wage=[1,8.25)} 0.1254682 0.8815789  0.1423221 1.651801 67
[7]  {age=[18,25.7)}                                    => {wage=[1,8.25)} 0.1498127 0.8791209  0.1704120 1.647195 80
[8]  {union=0,occupation=4}                             => {wage=[1,8.25)} 0.1048689 0.8484848  0.1235955 1.589793 56
[9]  {education=[12,14),experience=[0,9.17),race=3}  => {wage=[1,8.25)} 0.1142322 0.8472222  0.1348315 1.587427 61
[10] {education=[12,14),experience=[0,9.17)}         => {wage=[1,8.25)} 0.1367041 0.8390805  0.1629213 1.572172 73
```

Figure 12. The pruned dataset shows a majority of rules have 2 conditions in the LHS, moderate to high confidence with a support ranging 0.01-0.24 and all lift is > 1.3 (all RHS pertain to wages)(model 4).