|  |  |
| --- | --- |
|  | **2017** |
|  | Resultat d'imatges de liu university  Carles Sans Fuentes |

|  |
| --- |
| **[main features affecting earnings on population]** |
| The project tries to identify the main features on the population to earn more than 50K dollars per year. Those features are identified using a very well-known rule based algorithm called a priori algorithm, based on confidence, support and lift |

Index

[1. MOTIVATION 2](#_Toc485287506)

[Background 2](#_Toc485287507)

[Scope 3](#_Toc485287508)

[2. DATASET EXPLANATION 3](#_Toc485287509)

[3. THE A PRIORI ALGORITHM 4](#_Toc485287510)

[4. DATASET CLEANING 5](#_Toc485287511)

[Removals 5](#_Toc485287512)

[Data modification 5](#_Toc485287513)

[5. TEST RESULTS 6](#_Toc485287514)

[Winning more than $50K 6](#_Toc485287515)

[Winning less than $50K 9](#_Toc485287516)

[6. CONCLUSIONS 12](#_Toc485287517)

[Evaluation 12](#_Toc485287518)

[Discussion and comparison 14](#_Toc485287519)

[Criticism 15](#_Toc485287520)

[Forward steps 15](#_Toc485287521)

[7. CODE 15](#_Toc485287522)

# MOTIVATION

Inequality in salaries due to inequality of opportunities and discrimination has always been one of the major issues that has concerned the population among different countries. There is always though (unless there is direct evidence from the discrimination) some background to reason upon the election of the best candidate for a position: experience, education and several other aspects (e.g. the physical appearance in an interview) define a lot of times why is one candidate chosen over another. Nevertheless, are there intrinsic variables that can make someone being discriminated indirectly or directly given some prejudices on the population? It is alleged and acknowledged by different countries, for instance, that there is clear evidence of discrimination for the female gender on their jobs; not only about the difference on salaries for a similar or equal position than a man, but rather also about the chances to get to a higher position given its capabilities and hence earning more money either in the public or private sector.

The usage of computers and its incremental power and speed during the last decades have been crucial in order to evaluate and assess objectively how large these issues are as well as several other ones (e.g. gender discrimination by origin or color). Thus, the evaluation of large amounts of data (called today Big Data) using algorithms that evaluate and assess information efficiently in almost no time has been able to show to the society the dimension of certain problems with clear proof to discuss and argument the subject that matters. Nevertheless, when the analysis of data is done, it is always crucial that when you run algorithms and establish models, all variables available should be included in the first instance for the purpose of seeking the maximum evidence and the true cause of the issue. Thus, it is important to be aware that information can be presented and shown from a lot of perspectives, and that some of them may lack of a good evaluation, the good data or a wrong model. Thus, it is always important to be able to show in a clear matter your findings making them reproducible. For this reason, I will be using a basic algorithm (the apriori) in this Data Mining project at the university of Linköping to establish rules that evaluate the main variables from the US census in 1994 that affect whether a person won more than $50K per year.

## Background

The median salary (the income from the person who was at the 50% of the whole population by salary) in the US in 1994 was of $31,522[[1]](#footnote-1).

## Scope

The scope of this project is not about learning how the algorithm works in depth neither to find new revolutionary conclusions but understanding how to use algorithms in real data sets, learning how to use an algorithm (at my choice) in a software and derive conclusions from it.

# DATASET EXPLANATION

The dataset used for this data mining a project is a well-known dataset called adults[[2]](#footnote-2), implemented in several softwares and evaluated with different algorithms. The dataset, extracted by Barry Becker, contains information about the 1994 Census Boreau database in the US. More specifically, it contains 48,842 observations from people with 15 attributes to take into consideration with its following classes:

1. ***Age***: continuous variable related to the observations.
2. ***Workclass***: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. ***Fnlwgt***: The number of people the census takers believe that observation represents.
4. ***Education***: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
5. ***Education***-num: continuous.
6. ***Marital***-***status***: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
7. ***Occupation***: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
8. ***Relationship***: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
9. ***Race***: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
10. ***Sex***: Female, Male.
11. ***Capital***-***gain***: continuous.
12. ***Capital***-***loss***: continuous.
13. ***Hours***-***per***-***week***: continuous.
14. ***Native***-***country***: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
15. ***Winning50K***: whether the observations win more than $50000 or not: >50K, <=50K.

The idea for this data set is to find relations and importance of variables to the variable of winning more than $50000 or not. Having that said, the dataset shows that the Probability for the label '>50K' is 24.78% and Probability for the label '<=50K’: 75.22% (once unknowns are cleaned up on the data base provided[[3]](#footnote-3))

# THE A PRIORI ALGORITHM

The apriori algorithm is a frequent items set algorithm used to find association rules over transactional databases. It is really used in areas such as the food industry or the banking industry in order to understand costumer behavior and elaborate bundles and reccomed products to clients. The working procedure of the algorithm is starting with individual item sets and then enlarging them under some constraints, being the most typical ones the follwoing ones:

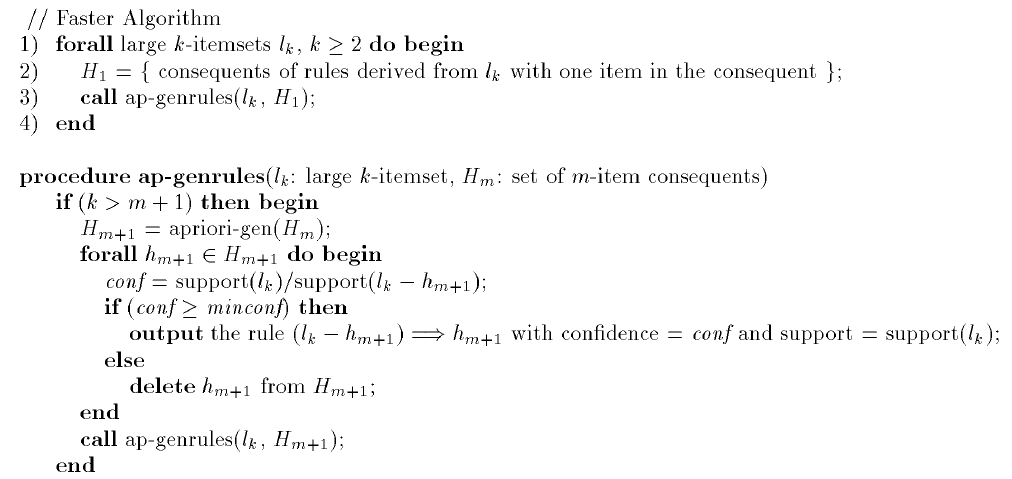
* *Minimum support*: Probability that a transaction contains product X and Y given all transactions.
* *Confidence*: Probablity of Y for all transactions containing X.
* *Lift*: measure that identifies the probability of a rule given the total possibility of that rule in the model. It is given by the formula lift (X) =Prob(X|Y) /Prob(X)
* *Target feature:* It is also typical to use some target features or products contained in the dataset as a constrain in order to assess rules based on this target feature.

Figure PseudoCode from the apriori algorithm in Han - Data Mining Concepts and Techniques 3rd Edition – 2012, page 253.

The pseudocode for the algorithm can be found in the following picture[[4]](#footnote-4):

One of the limitations of the algorithm is that it requires data to be discrete. Also, it can generate a huge amount of inefficient rules, so good constraints must be designed in order to be optimal.

# DATASET CLEANING

Since the apriori algorithm requires all data to be classified as categorical some modifications and variable elimination has been performed:

## Removals

The attribute *education-num* has been removed given that it provides the same information in then the categorical variable of *education*, which is already in bins.

## Data modification

To find a better set of large rules, variables which were continuous have been simplified in bins in the following way.:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Age*** | ***Fnlwgt*** | ***Capital gain*** | ***Capital loss*** | ***Hoursperweek*** |
| <30s | <100M | No Gain | No Loss | <20 hours |
| 30-40s | 100-200M | Gain | Loss | 20-34 hours |
| 40-50s | 200-300M |  |  | 35-44 hours |
| 50-65s | 300-400M |  |  | 45-60 hours |
| >65 | >400M |  |  | >60 hours |

*Age* has been simplified in decades mainly, being the last tram before retirement larger because I understand that larger changes on salary come before the 50s, so that the salary you have when you are 50 is already similar to your salary until you get retired.

*Fnlwgt* has been classified by how many people it concerns to, doing bins of 100.000 people.

*Capital gain* and *loss* has been classified as winnings, losses or no losses neither gains.

The *number of hours worked per week* has been classified as less than 20, considering this normally as almost sporadic work, 20-34 accounting for those people who work part time completely. From 35-44 it would stand for the amount of hours that a person does in an average job. From 45-60 counting for people who Works in more than one job or consultant, and then the variable more than 60.

# TEST RESULTS

The apriori algorithm used for our different evaluations belongs to the arules library from the R[[5]](#footnote-5) software. The methodology to find out rules has been as following:

* Separate the creation of rules from Winning more than 50K and less. In each case, it has been evaluated:
* The first 10 rules by confidence given lift higher than 1
* The first 10 rules by support given lift higher than 1
* The number of variables that appear more times in the rules to see which are those having more effect given some support, lift and confidence.
* Specific rules for Race = Black and Sex = Female

## Winning more than $50K

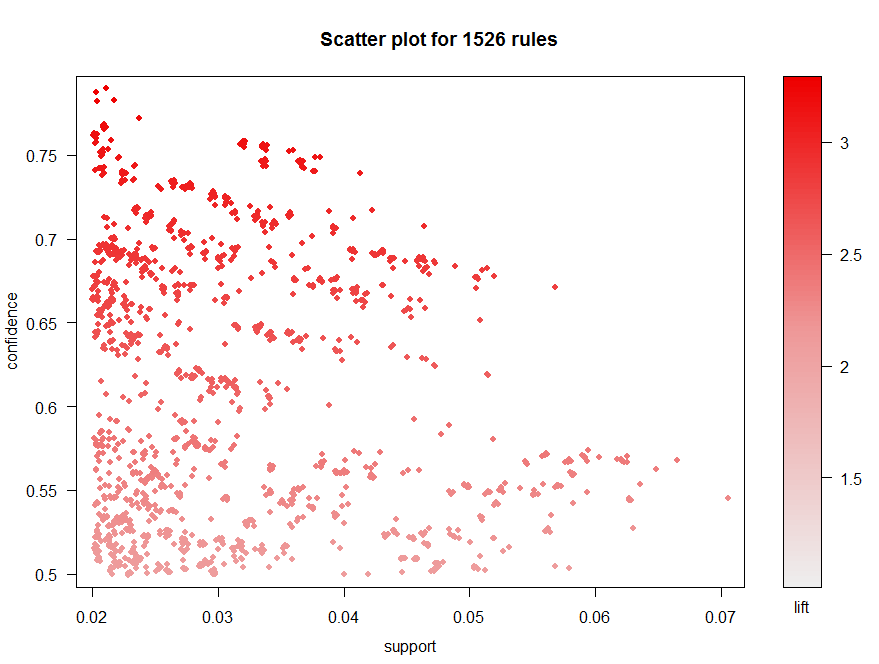
The rules created have as a result >50K on the right side (since it is what it is evaluated). The support and confidence that has been used to create the first basic set of rules have been 0.02 and 0.5. This can be seen as a little but not even 25% of data containing the variable >50, so 0.02 stands for a 10% rule to it, so that it is not that low. The Figure below shows the 1,526 rules created by support and confidence:

Figure 2 Scatterplot of 1526 rules created from the apriori algorithm with support and confidence equal to 0.02 and 0.5 respectively. Representation of the lift can also be seen in the right-hand side

Rules vary from support 0.02-0.07 and confidence 0.5-0.8. Nevertheless, rules with 50% of confidence might be low to extract good general rules, so more search is done by subsetting from the first rules only those with lift bigger than 1.05 and confidence equal to 0.7:

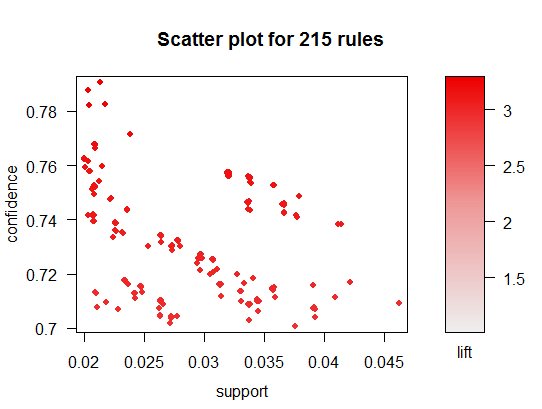


Figure 3 Scatterplot of 215 rules subsetted from the previous apriori algorithm general support and confidence constrained with lift and confidence equal to 1.05 and 0.7 respectively. Representation of the lift can also be seen in the right-hand side

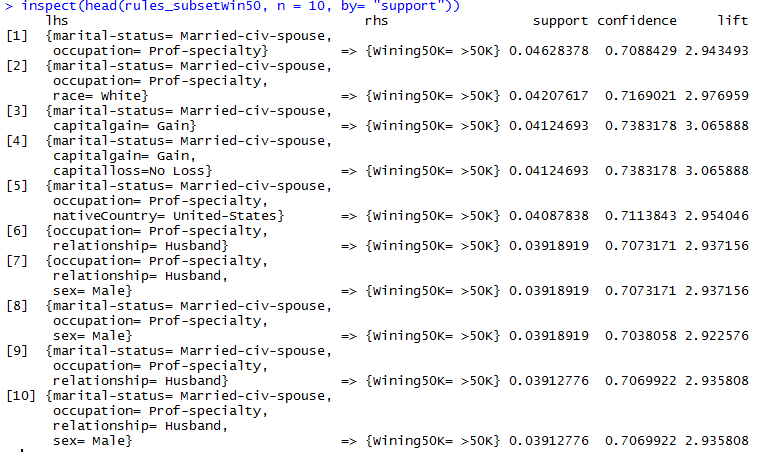
From this subset rules, here we can see the top 10 rules by confidence and by support:

Figure Top 10 rules by support

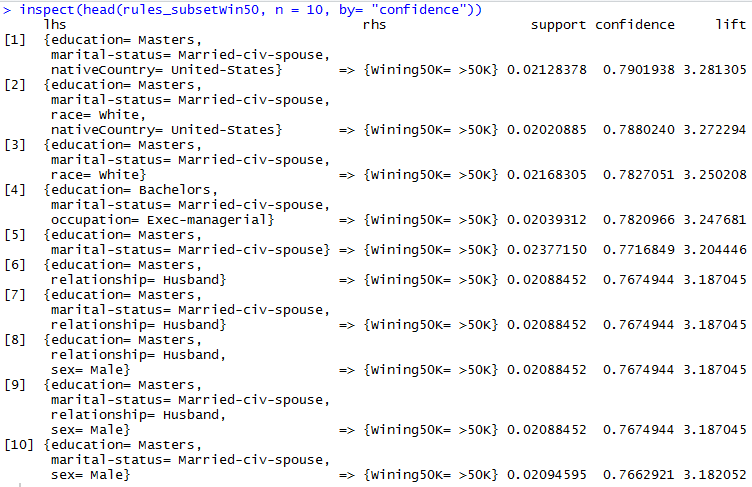


Figure Top 10 rules by confidence

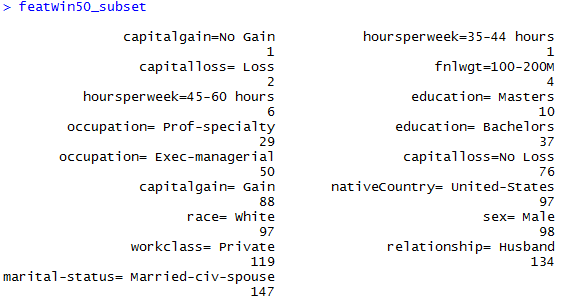
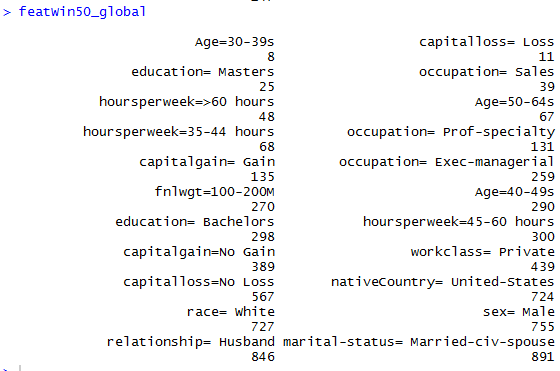
Also, it can be seen below the number of times the variables are repeated in the previous general rule set and subset:

Figure Count of the variables repeated in the set

Figure Count of the variables repeated in the subset

## Winning less than $50K

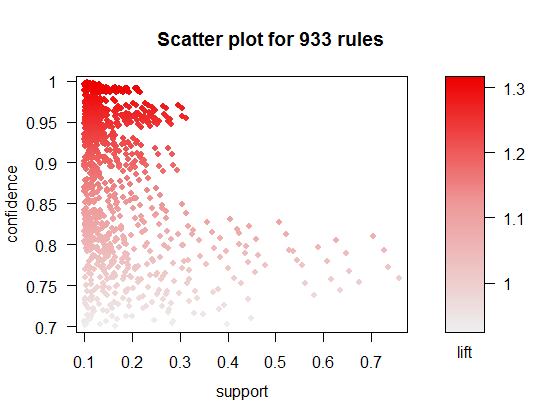
The rules have as a result the variable <50K on the right just (since it is what it is evaluated). The support and confidence that has been used to create the first basic set of rules have been 0.1 and 0.7. The Figure below shows 933 rules created by support and confidence:

Figure Scatterplot of 933 rules created from the apriori algorithm with support and confidence equal to 0.1 and 0.7 respectively. Representation of the lift can also be seen in the right-hand side

Rules vary from support 0.1-0.7 and confidence 0.7-1. More search is done by subsetting from the first rules only those with lift bigger than 1.05 and confidence equal to 0.7:

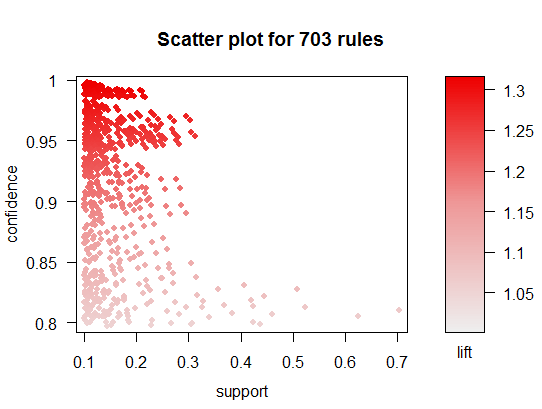


Figure Scatterplot of 703 rules subsetted from the previous apriori algorithm general support and confidence constrained with lift and confidence equal to 1.05 and 0.7 respectively. Representation of the lift can also be seen in the right-hand side

From this subset rules, here we can see the top 10 rules by confidence and by support:

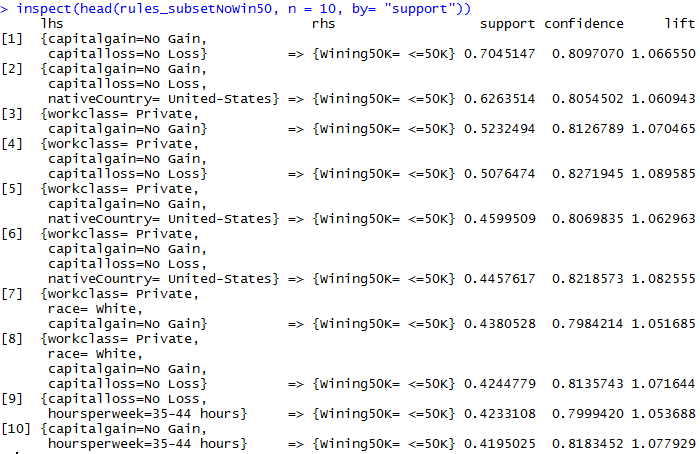


Figure Top 10 rules by support

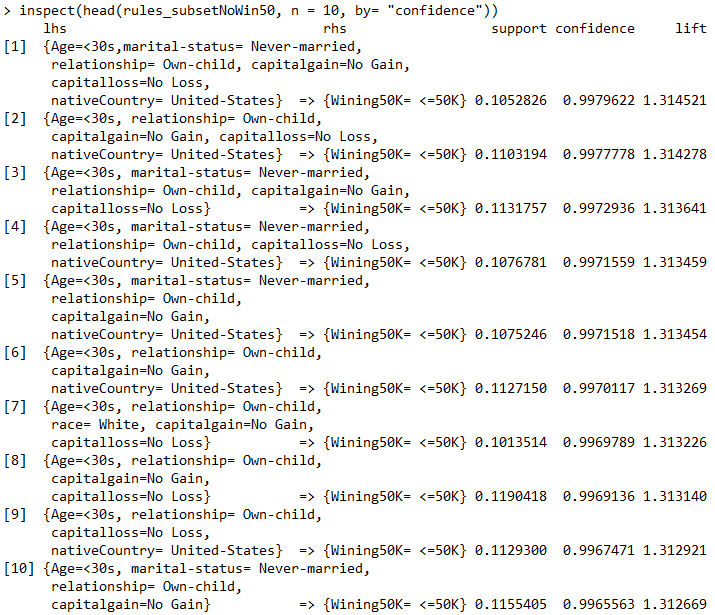


Figure Top 10 rules by confidence

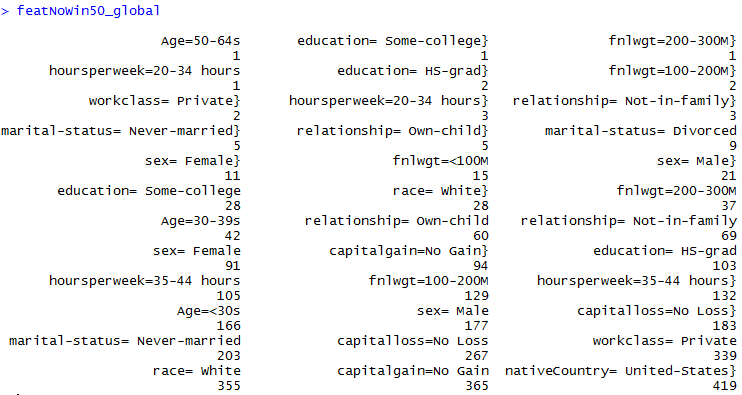
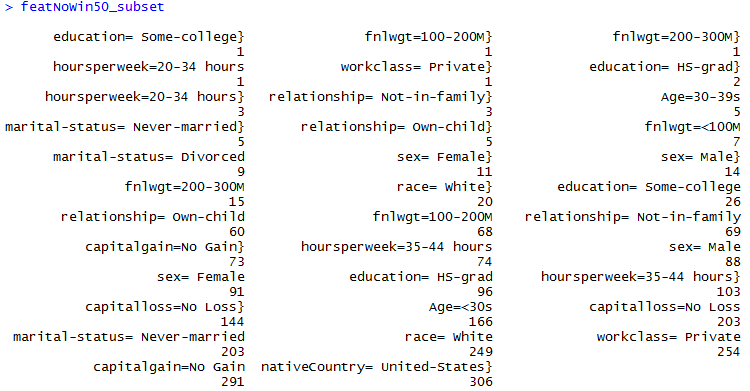
Also, it can be seen below the number of times the variables are repeated in the previous general rules set and subset:

Figure Count of the variables repeated in the subset

Figure Count of the variables repeated in the set

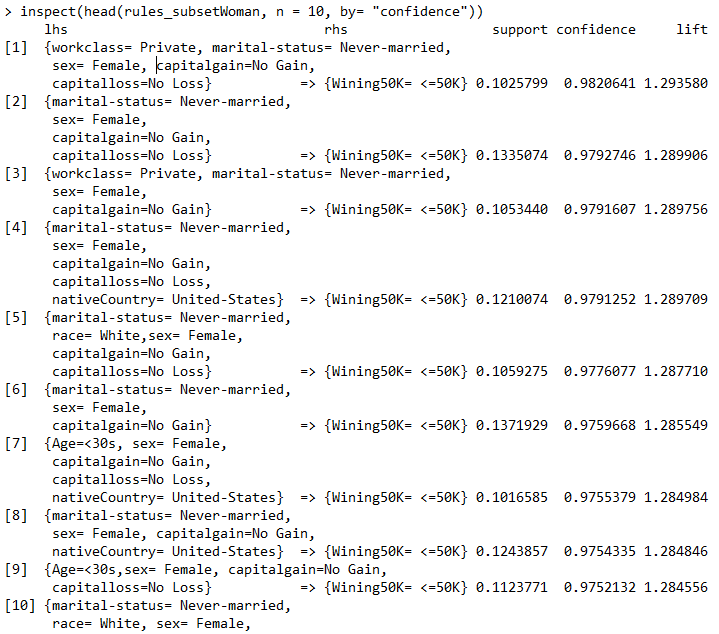


Figure Top 10 rules for woman winning <50K by confidence

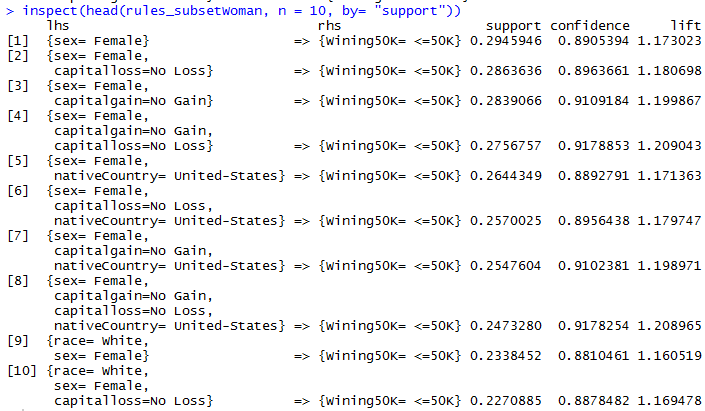


Figure Top 10 rules for woman winning <50K by support

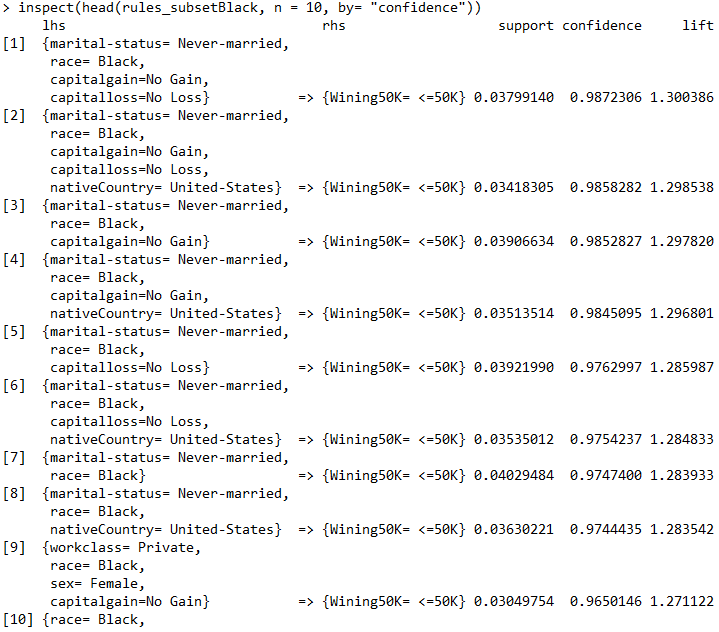


Figure Top 10 rules for race = Black winning <50K by confidence

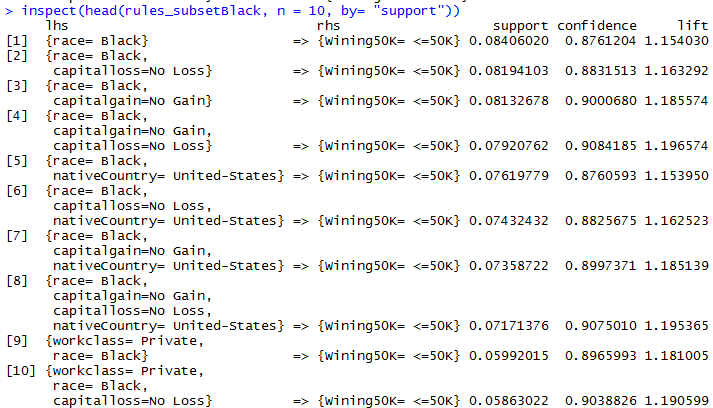


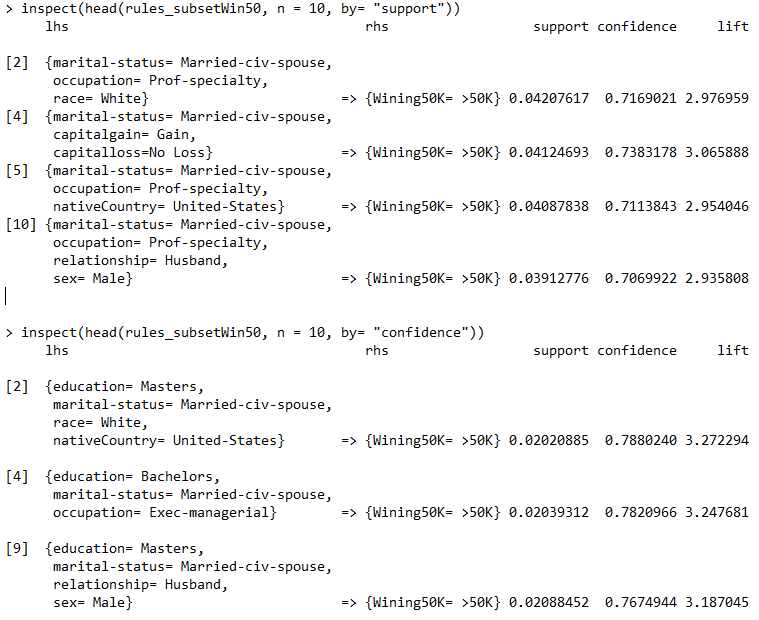
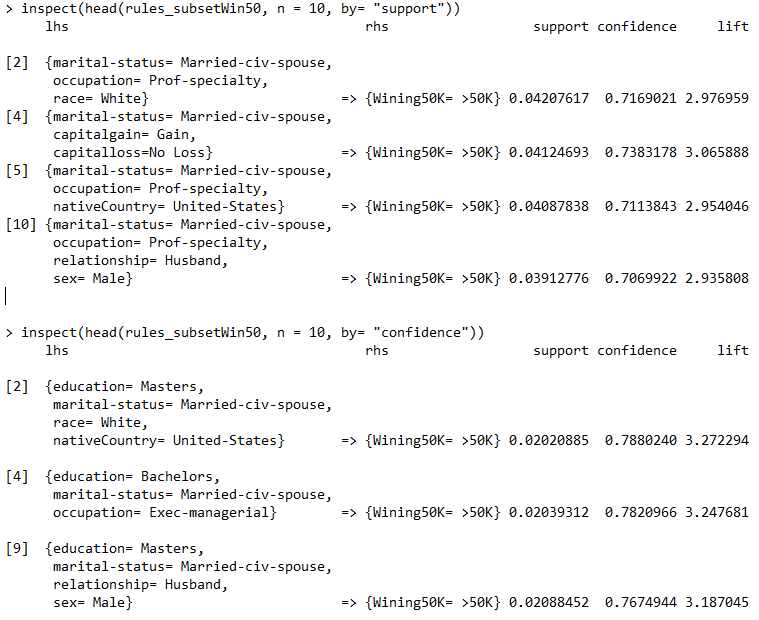
Figure Top 10 rules for race = Black winning <50K by support

# CONCLUSIONS

## Evaluation

After having cleaned the top 10 rules for winning more and less than $50K, the Figure below summarizes the best rules among the top 10 ones for the >$50K:

Figure Most important rules in the top 10 for winning >50K

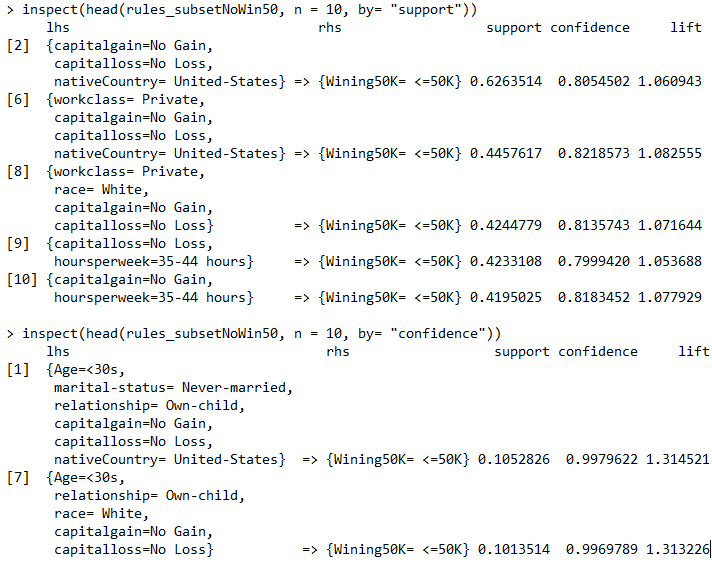
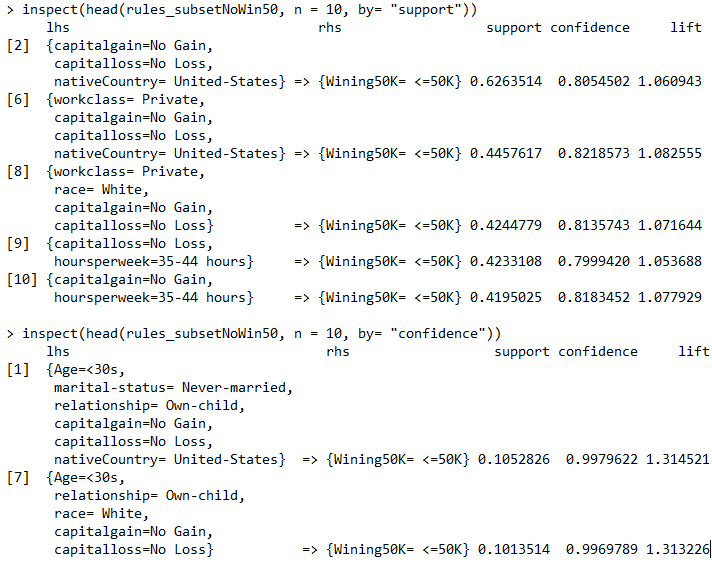


If you have a profesional specialty and you are husband, male or white, you are probable to win more than $50K. Moreover, if your education is master and you are white, husband or male, or if you have a bachelor’s diploma and work as an ex-managerial being married, your probabilities of also winning more than $50K are quite high. Also, the fact about having capital gains affect your probability of winning more than 50K.

The attributes that are more repeated for the rules on the set and subset are: being husband, married-civ-spouse, male, white, native-country equal to US, having capital gains, working on the private sector, being ex-managerial, or having at least bachelor and prof specialty and age 40-49.

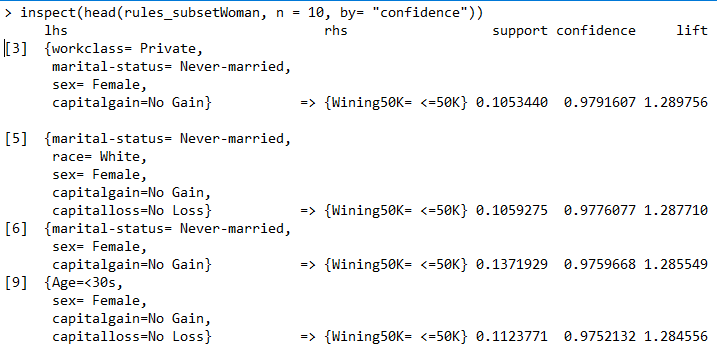
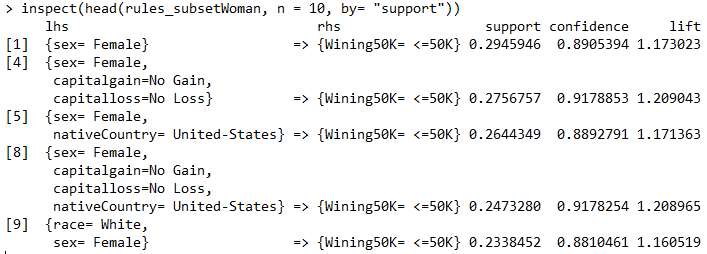
On the other hand, , if you don’t have capital gain or loss, if your working class is private and your working time per week is between 35-44 hours the probabilities of not winning $50K are higher. Also, if you have never been married and you are under 30 years old with a kid without capital gains or losses, then your probabilities of not winning $50K are almost 100%.

Figure Most important rules in the top 10 for winning <50K



The attributes that are more repeated for the rules on the set and subset are: being never-married or not in family, age<30, white, native country equal to US, no capital gains, sex = female, sex = male, education = HS-grad, working 35-44 hoursperweek.

Figure Most important rules from the top 10 for woman winning <50K



Given that woman are 0.33 of the cases, having 0.29 is about 88% of the woman winning less than 50K. The variable white seems to affect a lot as well as race, making the support lower just to 0.23 so 70%. Still, the biggest impact is the variable woman. It has been found that largest rules for woman earning more than 50K are being white and married being wife.

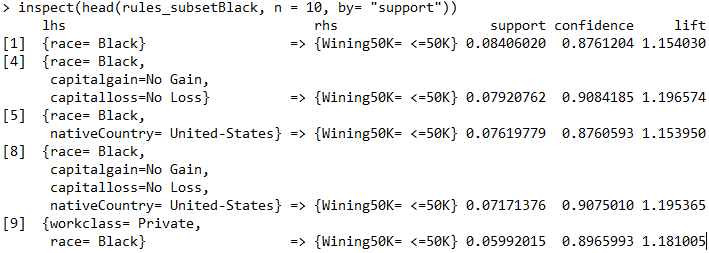
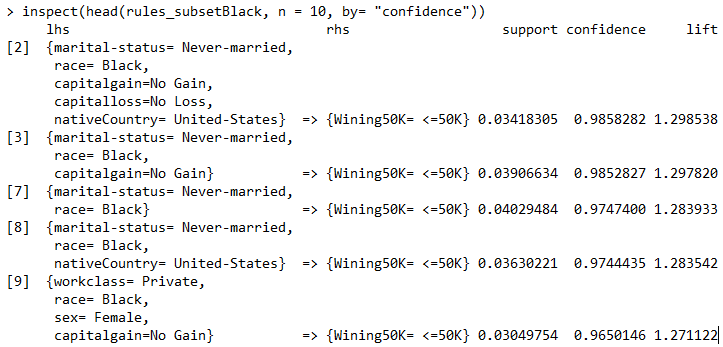


Figure Most important rules from the top 10 for race = Black winning <50K

Given that race = Black in the data set accounts for just 0.095 of the cases, having 0.084 earning less than $50K with 0.87 of confidence is quite impacting. One can say that this variable totally affects the probability of winnning more than $50K. No Loss, No Gain or native Country do not affect much (neither support or confidence change much). Also, if you are black and female, even if you work in the private sector, the probability of you winning less than $50K is 0.965. Again, if you are black and you have never been married, with support 0.363 and confidence 0.974 you will not be winning more than $50K. Also, being married= hubsand and race = black are the variables that affect more (with higher confidence and support) when winning more than $50K.

## Discussion and comparison

The main results got from the analysis are that variables such as being white, Husband or married with spouse with high education (at least bachelor) affect highly your chances of winning more than $50K, whereas the variables of being a woman, race = Black, not being married or being lower than 30 years affect almost void the chances of winning more than $50K.These results were expected given the high level of discrimination known in the US 1990s as well as the possibility for the major part of the population to study, biasing them towards having to work from little and leaving education. Also, as I have previously said, it seems to exist some concern for woman not winning more than $50K given that it is not easy to find rules with interesting variables winning more than $50K rather than being married, probably to someone who does win much more than 50K who has contacts for her to get a decent job. These results should make think to the US society about potential problems in the population and try to look for good explanations on them.

## Criticism

The apriori algorithm does generate too many rules a lot of times, some of them not being relevant neither useful or causal even if they look like so. Nevertheless, variables like being black and/or being woman as by itself seem to have a huge impact on population. It would be necessary though trying to find more concrete rules that could help classify and explain the issues on population better as well as getting the help of an expert in the topic to do more accurate searches by knowing what are we looking for.

## Forward steps

My work on analyzing this data set has been little, using just one algorithm to extract conclusions to see potential rules.It would be appropiate and necessary using other algorithms like C4.5, Naive-Bayes or Knn (like it has been done in orhter articles) in order to confirm better my evaluations and insights about the data set and so the US society in 1994.

# CODE

#install.packages("stringr")

#install.packages("plyr")

#install.packages("stats")

#install.packages("arules")

#install.packages("arulesViz")

#install.packages("stringr")

library(plyr)

library(stats)

library(arulesViz)

library(arules)

library(stringr)

###Reading data set

data<- read.csv("C:/Users/Carles/Desktop/Data mining/dataset.1.txt",sep =",", stringsAsFactors = FALSE)

class(data)

table(data[,1])

#####################################################################################

####################################CLEANINGDATA##################################### #############################################################

##############################################################################################################

colnames(data)<-c("Age", "workclass", "fnlwgt", "education","education-num", "marital-status", "occupation", "relationship", "race", "sex","capitalgain", "capitalloss", "hoursperweek", "nativeCountry", "Wining50K")

##Droping education num because it stands for the same as education but continuous

drops<- c("education-num")

data<- data[ , !(names(data) %in% drops)]

table(data[["Wining50K"]])

### <=50K >50K

### 24719 7841

##wining +50K == to 1, less than 50k equals to 0

levels(data[["Wining50K"]])<-c(0,1)

#Data does not contain NA

apply(data, 2, anyNA)

#################################APRIORI #################################################

#install.packages("arules")

#colnames(data)<-c("Age", "workclass", "fnlwgt", "education","education-num", "marital-status", "occupation", "relationship", "race", "sex","capitalgain", "capitalloss", "hoursperweek", "nativeCountry", "Wining50K")

min(data[,1]); max(data[,1]);

hist(data[,1])

thirties<-which(data[,1]<30)

fourties<-which((data[,1]>=30)&(data[,1]<40))

fifties<-which((data[,1]>=40)&(data[,1]<50))

sixties<-which((data[,1]>=50)&(data[,1]<65))

pensioners<- which(data[,1]>=65)

data[thirties,1]<-"<30s"

data[fourties,1]<-"30-39s"

data[fifties,1]<-"40-49s"

data[sixties,1]<-"50-64s"

data[pensioners,1]<-">65"

table(data[,1])

### <30s >65 30-39s 40-49s 50-64s

### 9711 1336 8612 7175 5726

colnames(data)[c(3,5,11,13)]##[1] "fnlwgt" "education-num" "capitalgain" "hoursperweek"

#### Final weight: The \# of people the census takers believe that observation represents. We will be ignoring this variable

hist(data[,3])

min(data[,3]); max(data[,3]);

belowhundred<- which(data[,3]<100000)

hundred<- which((data[,3]>=100000)&(data[,3]<200000))

twohundred<- which((data[,3]>=200000)&(data[,3]<300000))

threehundred<-which((data[,3]>=300000)&(data[,3]<400000))

fourhundred<- which(data[,3]>=400000)

data[belowhundred,3]<-"<100M"

data[hundred,3]<-"100-200M"

data[twohundred,3]<-"200-300M"

data[threehundred,3]<-"300-400M"

data[fourhundred,3]<-">400M"

table(data[,3])

class(data[,3])

#### "capital gain"

hist(data[,10])

min(data[,10]); max(data[,10]);

NoGain<-which(data[,10] == 0)

Gain<-which(data[,10]>1)

data[NoGain,10]<-"No Gain"

data[Gain,10]<-" Gain"

table(data[,10])

#### "capital loss"

hist(data[,11])

min(data[,11]); max(data[,11]);

NoLoss<- which(data[,11]==0)

Loss<- which(data[,11]>0)

data[NoLoss,11]<-"No Loss"

data[Loss,11]<-" Loss"

table(data[,11])

#### "hoursperweek"

min(data[,12]); max(data[,12]);

hist(data[,12])

lesstwenty<- which(data[,12]<20)

twenty<- which((data[,12]>=20)&(data[,12]<35))

thirtyfive<-which((data[,12]>=35)&(data[,12]<45))

fortyfive<-which((data[,12]>=45)&(data[,12]<60))

PlusSixty<-which(data[,12]>=60)

data[lesstwenty,12]<-"<20 hours"

data[twenty,12]<-"20-34 hours"

data[thirtyfive,12]<-"35-44 hours"

data[fortyfive,12]<-"45-60 hours"

data[PlusSixty,12]<-">60 hours"

table(data[,12])

Myframe<- as.data.frame(data)

DataFactor<-Myframe

for(i in 1:ncol(data)){

#levels(data[,i])<-names(table(data[,i]))

DataFactor[,i]<- as.factor(data[,i])

}

##############################################################################################################

####################################A PRIORI ALGORITHM#############################################################

##############################################################################################################

####MAIN FUNCTIONS TO CHECK FROM THE RULES STATED, which ones are the repeated variables most variables

MainWordsList<-function(data= resultswin50\_global , start=2, stop=22, numbercommas=6){

require(stringr)

####just taking into account those which has more than 5 variables, counted as comma separated

data<-data[which(str\_count(data[,1], ',')>=numbercommas),]

data[,1]<- as.character(data[,1])

for(i in 1:length(data[,1])){

##Getting rid of last variable and first variable for counting most important features

data[i,1]<- substr(data[i,1],start,nchar(data[i,1])-stop)

}

featWin50<-table(unlist(strsplit(as.character(data[,1]), split = c(","))))

return(featWin50)

}

##### RULES FOR WIN +50K###################

Mytransmatrix <- as(DataFactor, "transactions")

Win50<- apriori(Mytransmatrix, parameter = list(supp = 0.02, conf = 0.5, target = "rules"),

appearance = list(rhs = c("Wining50K= >50K"),

default="lhs"))

plot(Win50, method = NULL, measure = "support", shading = "lift",

interactive = FALSE, data = NULL, control = NULL)

rules\_subsetWin50 <- subset(Win50, subset=(lift>1.05& confidence>0.7))

plot(rules\_subsetWin50, method = NULL, measure = "support", shading = "lift",

interactive = FALSE, data = NULL, control = NULL)

##larger by support

inspect(head(rules\_subsetWin50, n = 10, by= "support"))

##larger by confidence

inspect(head(rules\_subsetWin50, n = 10, by= "confidence"))

resultswin50\_global<- as(Win50, "data.frame");

resultswin50\_subset<- as(rules\_subsetWin50, "data.frame");

featWin50\_subset<- sort(MainWordsList(resultswin50\_subset,2,22,1))

featWin50\_global <- sort(MainWordsList(resultswin50\_global,2,22,1))

##### RULES FOR WIN -50K###################

NoWin50<- apriori(Mytransmatrix, parameter = list(supp = 0.10, conf = 0.7, target = "rules"),

appearance = list(rhs = c("Wining50K= <=50K"),

default="lhs"))

plot(NoWin50, method = NULL, measure = "support", shading = "lift",

interactive = FALSE, data = NULL, control = NULL)

rules\_subsetNoWin50 <- subset(NoWin50, subset=(lift>1.05& confidence>0.7))

plot(rules\_subsetNoWin50, method = NULL, measure = "support", shading = "lift",

interactive = FALSE, data = NULL, control = NULL)

##larger by support

inspect(head(rules\_subsetNoWin50, n = 10, by= "support"))

##larger by confidence

inspect(head(rules\_subsetNoWin50, n = 10, by= "confidence"))

resultsNowin50\_global<- as(NoWin50, "data.frame");

resultsNowin50\_subset<- as(rules\_subsetNoWin50, "data.frame");

dim(resultsNowin50\_subset)

featNoWin50\_subset<- sort(MainWordsList(resultsNowin50\_subset,2,22,1))

featNoWin50\_global<- sort(MainWordsList(resultsNowin50\_global,2,22,1))

table(data[,8])

##### RULES FOR WIN -50K###################

###Rules for female

NoWin50Female<- apriori(Mytransmatrix, parameter = list(supp = 0.10, conf = 0.7, target = "rules"),

appearance = list(rhs = c("Wining50K= <=50K"),

items = c("sex= Female"),

default="lhs"))

rules\_subsetWoman <- subset(NoWin50Female, (lhs %in% "sex= Female"))

inspect(head(rules\_subsetWoman, n = 10, by= "confidence"))

inspect(head(rules\_subsetWoman, n = 10, by= "support"))

table(data[,9])[1]/sum(table(data[,9]))

### Female

### 0.3308047

###Rules for female to win 50K

Win50Female<- apriori(Mytransmatrix, parameter = list(supp = 0.02, conf = 0.2, target = "rules"),

appearance = list(rhs = c("Wining50K= >50K"),

items = c("sex= Female"),

default="lhs"))

rules\_subsetWomanwin50 <- subset(Win50Female, (lhs %in% "sex= Female"))

inspect(head(rules\_subsetWomanwin50, n = 10, by= "confidence"))

inspect(head(rules\_subsetWomanwin50, n = 10, by= "support"))

table(data[,9])[1]/sum(table(data[,9]))

### Female

### 0.3308047

###Rules for race black

NoWin50Black <-apriori(Mytransmatrix, parameter = list(supp = 0.03, conf = 0.5, target = "rules"),

appearance = list(rhs = c("Wining50K= <=50K"),

items = c("race= Black"),

default="lhs"))

rules\_subsetBlack <- subset(NoWin50Black, (lhs %in% "race= Black"))

inspect(head(rules\_subsetBlack, n = 10, by= "confidence"))

inspect(head(rules\_subsetBlack, n = 10, by= "support"))

table(data[,8])[3]/sum(table(data[,8]))

### Black

####0.09594595

###Rules for race black to win 50K

Win50Black <-apriori(Mytransmatrix, parameter = list(supp = 0.005, conf = 0.2, target = "rules"),

appearance = list(rhs = c("Wining50K= >50K"),

items = c("race= Black"),

default="lhs"))

rules\_subsetBlackwin <- subset(Win50Black, (lhs %in% "race= Black"))

inspect(head(rules\_subsetBlackwin, n = 10, by= "confidence"))

inspect(head(rules\_subsetBlackwin, n = 10, by= "support"))

table(data[,8])[3]/sum(table(data[,8]))

### Black

####0.09594595

1. Reference to this information can be found in <http://www.davemanuel.com/median-household-income.php> [↑](#footnote-ref-1)
2. More information about the data set as well as the possibility of downloading it can be found in <https://archive.ics.uci.edu/ml/datasets/adult> [↑](#footnote-ref-2)
3. Evaluation and information just explained is found in <http://www.cs.toronto.edu/~delve/data/adult/adultDetail.html> [↑](#footnote-ref-3)
4. Since it is an algorithm explained in class, only the main idea of it is given [↑](#footnote-ref-4)
5. The library arules can be found in <https://cran.r-project.org/web/packages/arules/arules.pdf> [↑](#footnote-ref-5)