Final Project: Association Rules

Carlos Manuel Sequí Sánchez

12 de febrero de 2019

Dataset description.

I have chosen Suicide Rates in Germany dataset stored in R datasets. Data from Heuer(1979) on suicide rates in West Germany classified by age, sex, and method of suicide. The dataset is made up of 306 instances of 6 attributes each one, including the frequency with which each type of suicide has occurred.

Attributes:
Freq: frequency of commited suicides with the same attributes.
sex: factor indicating sex (male, female).
method: factor indicating method used (cookgas, drown, gun, hang, jump, knife, other, poison, toxicgas).
age: age (rounded).
age.group: factor. Age classified into 5 groups.
method2: factor indicating method used (same as method but cookgas and toxicgas are merged with gas).

Loading the dataset.

```
data = read.csv("Suicide.csv")
```

Preprocessing data.

Once we have read the data, let's make some interesting changes on them that will make us easier to analize them. The objective is to remove redundant and unnecessary information, such as id, suicide instances that never occurred (those which have Freq attribute equal to zero) and age and method columns, which are redundant with age.group and method2 attributes respectively.

In order to work correctly with our dataset, we are going to apply the frequency attribute into our dataset, replicating instances so many times as their Freq attribute indicates.

After this, let's have a look to a firstly explanation of our dataset:

```
##      sex        age.group         method
##  female:19363   10-20: 5768   hang   :20377
##  male  :33819   25-35:11945   poison :17565
##                 40-50:14023   other  : 3284
##                 55-65:12555   gun    : 3118
##                 70-90: 8891   jump   : 2845
##                               drown  : 2649
##                               (Other): 3344
```

We get a little idea of the major classes of each attribute:

- **Sex:** male
- **Age group:** 40-50
- **Method:** hang and poison

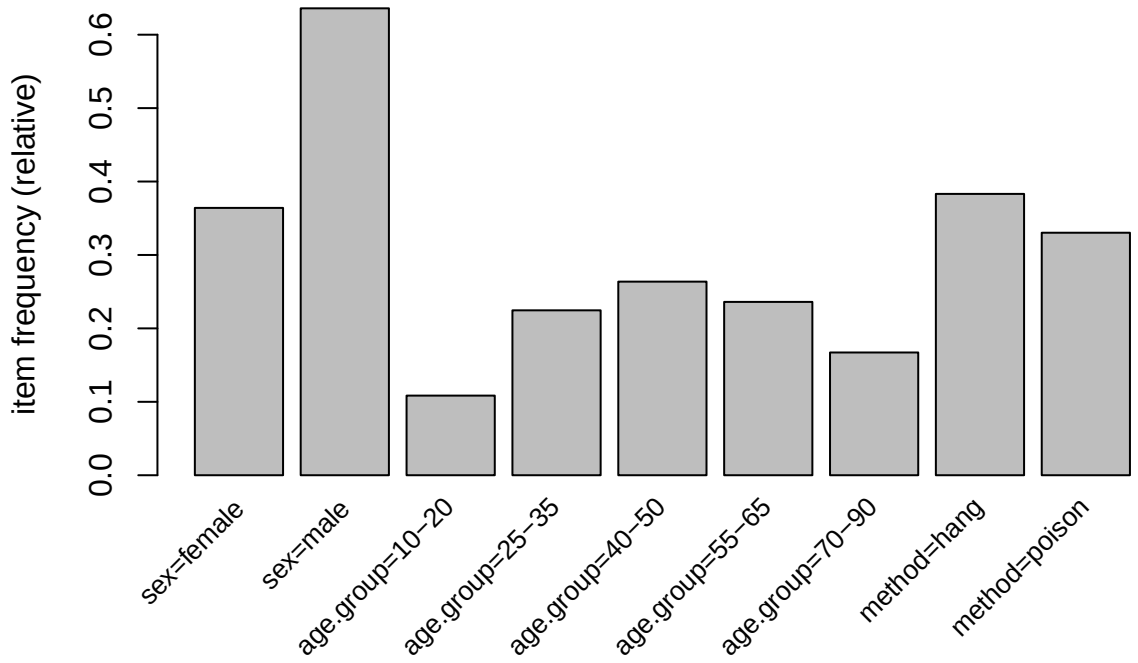
Once this is done, we only have 3 attributes, of which 2 we are going to transform into denied items: sex and age.group.

Now we have the same 3 attributes but, 2 of them splitted into different attributes. Sex attribute has been separated into male or female, and age.group into teenager (10-20), young(25-35), middle-aged(40-50), senior(55-65) and old(70-90).

Before converting our data.frame in transactions to extract association rules, we transform each attribute into a factor and then create transactions.

Analyzing data.

And then we are ready to display the item frequency plot (using non-denied items transactions in order to display clearer information):



Considering the frequency plot we can extract some information from our dataset:

- the most part of the registered suicides in West Germany were from middle aged people between 40-50 years old (27
- the most part of the population who decided to do it were men (64
- there are very few cases of people who commited suicide with: gas, knife, jump, drown, gun and other. All of these kind of suicides have a support lower than 0.1 in this dataset (less instances than 5300). The rest of the cases were suicides with hang and poison methods.

Now that we are sure of some patterns of sex, age and suicide method thanks to our first analysis, we're going to use transactions in order to extract rules based on a specific value of some classic measures such as

support, confidence and lift (using denied items transactions interpretation). Having this idea as target, let's firstly, using Apriori's method, try to find rules with support values higher than 0.1 and confidence values higher than 0.8.
Once we have this set of rules, we select the ones which have support values lower than 0.4 to extract not obvious information, as well as different lift values (lower or higher than 1, never equal to 1, which means total independency of items).

```
# first we apply our initial conditions with Apriori
rules <- apriori(transactions, parameter = list(support = 0.1, confidence = 0.8, minlen = 2))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.8    0.1    1 none FALSE            TRUE       5     0.1      2
##  maxlen target   ext
##      10  rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 5318
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[22 item(s), 53182 transaction(s)] done [0.03s].
## sorting and recoding items ... [16 item(s)] done [0.00s].
## creating transaction tree ... done [0.01s].
## checking subsets of size 1 2 3 4 5 6 7 done [0.00s].
## writing ... [1155 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
# and then we are ready to extract interesting rules
```

Considering the rules pruned with support values lower than 0.4, confidence lower than 1, and lift higher than 1, we can assume this information immediately extracted:

- 1.people who used poison as suicide method, were not old people.
- 2.women who commited suicide weren't neither teenager nor young.
- 3.people who commited hang suicide weren't neither teenager nor young.
- 4.men who commited poison suicide were not senior.

And watching these rules, we can deduce some more ideas:

- men who commited poison suicide were teenager, young or middle aged men (rules 1 and 4)
- most women who commited poison suicide where middle aged and senior women (rules 1 and 2)

Let's investigate some more about the topic changing the desired lift value (now we are looking for lift values < 1):

- 1.men who commited hang suicide were not old.
- 2.people who commited hang suicide were neither teenager nor young nor old.
- 3.teenagers commonly did not prefer to commit suicide using poison method
- 4.women who commmited poison suicide were not mostly old

A little more digging to find some more clues could be search for men who commited poison suicide:

- 1.men who commited poison suicide were not senior
- 2.men who commited poison suicide were not old
- 3.men who commited poison suicide were not teenager

Now we analyze specific ideas: In order not to focus the analysis in the attributes with major support (such as male=TRUE, method=hang or age=middleAged), I'm going to extract some new information directly from the dataset.

Firstly, for example, let's see the preferred suicide method of each age range:

```
# TEENAGERS
names(which(table(data[data$teenager,]$method) == max(table(data[data$teenager,]$method))))
```

```
## [1] "poison"
```

```
# YOUNG PEOPLE
names(which(table(data[data$young,]$method) == max(table(data[data$young,]$method))))
```

```
## [1] "poison"
```

```
# MIDDLE AGED PEOPLE
names(which(table(data[data$middleAged,]$method) == max(table(data[data$middleAged,]$method))))
```

```
## [1] "hang"
```

```
# SENIOR PEOPLE
names(which(table(data[data$senior,]$method) == max(table(data[data$senior,]$method))))
```

```
## [1] "hang"
```

```
# OLD PEOPLE
names(which(table(data[data$old,]$method) == max(table(data[data$old,]$method))))
```

```
## [1] "hang"
```

It seems that younger people preferred poison suicide method instead of hang. Now we're removing hang and poison methods to see which methods were mostly used (after those two) in each age group:

```
# TEENAGERS
names(which(table(data[data$teenager,]$method) == max(table(data[data$teenager
                                & !(data$method=="hang" | data$method=="poison"),]$method))))
```

```
## [1] "other"
```

```
# YOUNG PEOPLE
names(which(table(data[data$young,]$method) == max(table(data[data$young
                                & !(data$method=="hang" | data$method=="poison"),]$method))))
```

```
## [1] "other"
```

```
# MIDDLE AGED PEOPLE
names(which(table(data[data$middleAged,]$method) == max(table(data[data$middleAged
                                & !(data$method=="hang" | data$method=="poison"),]$method))))
```

```
## [1] "gun"
```

```
# SENIOR PEOPLE
names(which(table(data[data$senior,]$method) == max(table(data[data$senior
                               & !(data$method=="hang" | data$method=="poison"),]$method))))
```

```
## [1] "drown"
```

```
# OLD PEOPLE
names(which(table(data[data$old,]$method) == max(table(data[data$old
                                & !(data$method=="hang" | data$method=="poison"),]$method))))
```

```
## [1] "drown"
```

After hang and posion methods:

- Young people used non registered suicide methods.
- Middle aged people used gun suicide method, maybe because of their capability to get a weapon.
- Older people preferred drown suicide method, maybe because this is an easy method.

Let's see now for instance, the proportion between male and female teenagers that commit suicide:

```
dim(data[data$male & data$teenager,])[1]
```

```
## [1] 4298
```

```
dim(data[data$female & data$teenager,])[1]
```

```
## [1] 1470
```

Now we focus on the method of suicide taken by this age group (teenagers, sipplitted by sex):

```
# hang
dim(data[data$male & data$teenager & data$method == "hang",])[1]
```

```
## [1] 1524
```

```
dim(data[data$female & data$teenager & data$method == "hang",])[1]
```

```
## [1] 212
```

```
# poison
dim(data[data$male & data$teenager & data$method == "poison",])[1]
```

```
## [1] 1160
```

```
dim(data[data$female & data$teenager & data$method == "poison",])[1]
```

```
## [1] 921
```

```
#jump
dim(data[data$male & data$teenager & data$method == "jump",])[1]
```

```
## [1] 189
```

```
dim(data[data$female & data$teenager & data$method == "jump",])[1]
```

```
## [1] 131
```

```
# gun
dim(data[data$male & data$teenager & data$method == "gun",])[1]
```

```
## [1] 512
```

```
dim(data[data$female & data$teenager & data$method == "gun",])[1]
```

```
## [1] 25
```

```
# drown
dim(data[data$male & data$teenager & data$method == "drown",])[1]
```

```
## [1] 67
```

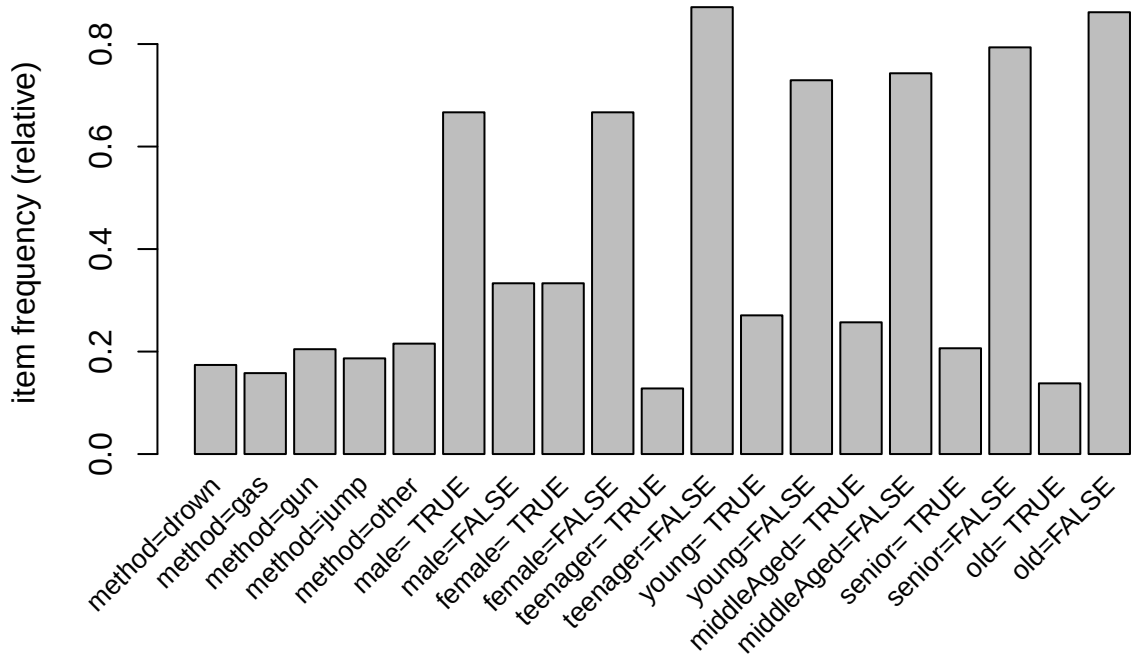
```
dim(data[data$female & data$teenager & data$method == "drown",])[1]
```

```
## [1] 30
```

As in the whole dataset, girls usually preferred poison suicide methods instead of hang methods, unlike men, who preferred hang method. It seems too that the proportion between men and women is the same in each age group of the dataset.

Let's create now a new dataset removing instances that used hang and poison methods to analyze new behaviors in our data:

```
newData = data[data$method!="hang" & data$method!="poison",]
```



new information extracted **when removing poison and hang suicide methods**:

- the difference between the amount of male and female amount of suicides increases (majority for male)
- the most used method is "other"" (unregistered methods) commited, as we concluded before, by young people (which is the age range with the highest suicide rate)

Analysis of rules by groups.

To extract stronger rules we try to find rules which accomplish the next requirements:

- (A->B) with (A -> !B). I can't find rules that meet these requirements.
- (A->B) or (!A -> !B). For instance: (male & middleAged -> hang | !male & !middleAged -> !hang)
- (A,B->!C). For instance: (male & hang -> !teenager)

Trying to extract information from our set of rules by analyzing groups of rules in the way indicated, we realize that we are only reaching redundant and unnecessary rules, which don't contribute to make us think about new ideas to investigate or new information to extract, so we conclude that data analysis by group of rules is not interesting in this case.

Final hypothesis based on analyzed data.

As we can observe both in item frequency plot and in the specific analysis of rules, in terms of the age of suicide, we arrive at the conclusion that the age group most affected by this event is that of people between 40 and 50 years old (middle aged people) where, in addition, approximately 65% of the population is male, as we have seen. Due to the date of obtaining the data associated with the dataset (1979) and assuming that the data may have been collected between 1970 and 1979, We can base our main hypothesis on the fact that this type of suicide is so common (men between 40 and 50 years old) due to the long period of tension in both western and eastern Germany (although the dataset refers exclusively to the western area) during years of the Second World War (1939-1945), and the outbreak of the Cold War (1947-1989) after the postwar period. We can think that both age and gender can be associated with people who have suffered the aftermath of World War II from when they were only between 15 and 25 years old, having participated in her (especially men) or just living her adolescence in that period of tension. In addition, as I have already mentioned, after the Second World War, the Cold War arrived, in which the German country was divided into 2 by the German Federal Republic (USA, Great Britain and France) and the German Democratic Republic (USSR). To these events we can add the terrible creation of the Berlin Wall in 1961 by the Soviets in order to eradicate the flight of German refugees from the eastern part of the country to the western part, which generated an increase in the already accumulated tension in the society of the time in that country. In conclusion, it is possible that the large number of suicides committed in this era are due to the psychological consequences of having lived through a World War and an endless period of tension with the Cold War.

The next two generations in terms of number of suicides are just adjacent to middle aged people, who lived situations similar to them but at different stages of their lives:

- On the one hand there is the senior group (55-65): of which most of them participated actively in the Second World War (they were between 25 and 35 years old at the time).
- On the other hand is the young group (25-35): who lived the Second World War in the first stages of their lives (between 5 and 15 years), which are probably the most important in the psychological development of the people.

From my point of view, these three generations (from 25 to 65 years old) are the most aware of the reality that occurs in society, that is, people under 20 and over 50 see life from another point of very different sight, from which surely the feeling of overwhelm by the events of the time does not produce stress enough to commit suicide. Because of this, generations of teenagers (10-25) and older people (> 70) do not commit as many suicides as other age groups.

Another possible conjecture that can cover a certain part of the suicides collected in the dataset is the existence of reprisals taken against Nazi thought in the post-war period, that is, people who do not want to continue living in post-war society with the feeling of failure and the persecution under punishment of their Nazi ideals both in the Federal Republic of Germany (West) and in the German Democratic Republic (East). For example, in 1945, it is known that in Demmin (although it was from the northeast of Germany) between 700 and 1000 people committed suicide before the arrival of the Red Army. It can even be thought that, given the time and the territory, it may be that many of the suicides produced were committed in a compulsory manner by the mandate of the political justice established in the country at that time.

On the other hand we can see how most of the suicides come from the hang method and from the poison method (higher being the hang one on the part of the men and the poison one being higher on the part of the women). This may be due to the desire not to be noisy when making a decision like this, instead of choosing methods such as gun or jump, to prevent everyone from realizing the action committed. In addition, the hang method is slightly more widely used than the poison method in general, perhaps due to the reliability of

the method used or simply the possibility of achieving an easy method that offers little suffering to commit suicide with poison. This type of behavior with respect to the type of suicide is met at any age.