Itemsets and Association Rules Mining
Practical work
Maria Afara
M2 TAL

Prepare the data for SPMF

Each multivalued attribute is transformed into single-valued attributes which are then associated to a number,in order to do that, I have generated a dictionary in this form:
Given a list of the attributes I want to keep:
For example **cols=["ETUD","INAT","SEXE"]**
A 2 dimensional dictionary is created which have as a key, the attribute name and as a value, another dictionary of which each key represents a unique value of its key attribute with a corresponding unique integer as a value in the second dictionary.
For example, the generated dictionary of the above list of attributes is :

```
{'ETUD': {'2': 0, '1': 1},
 'INAT': {'11': 2, '12': 3, '21': 4},
 'SEXE': {'1': 5, '2': 6}}
```

The **ETUD** attribute is multivalued i.e it may have several values, so each value corresponds to a unique integer (incremental integer).
The values 2 and 1 respectively corresponds to Non and Oui according to the tables provided in the **SignificationDesVariables.pdf** file, i.e whether educated or not.
Same does for the attributd **INAT** and **SEXE** and even all the attributes in the dataset.

Then for each transaction in the dataset a mapping was created from the unique items in the dataset to integers so that each item corresponds to a unique integer according to the dictionary created.
For example having a transaction as the following:

```
"ETUD","INAT","SEXE"
  1   ,  11      ,  1
```

It will be transformed to : **'1 2 5'** which means that the individual that is represented by this transaction have the value 1 for attribute **ETUD** , and the value 11 for **INAT** (which is represented by 2) and have the value 1 for **SEXE** attribute which represented by 5
And in order to interpret what each value means we have to check again the

SignificationDesVariables.pdf file.

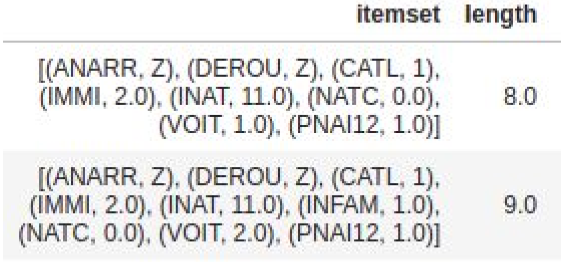
Decoder:
A reverse mapping was created from the integers to the items, so that the item names with its corresponding value could be written in the final output file.
And in order to do that I have also created the reverse dictionary for decoding at the same time while creating the encoder dictionary.
For example:
Having the following itemset
"1 5 #SUP: 53"
Its decoding will be **"('ETUD', '1') ('SEXE', '1') #SUP: 53"**

StoryTelling

The best storytelling with data focuses on a concise narrative, without including extra information that can distract from the points you're trying to make. That means you've got to go digging for *just* the right data, because often the best nuggets are hidden beneath the top-line results.
As for this survey me as a journalist I am interested in studying the combination of some itemsets, not arbitrary item sets which contain some attributes in which may lead to some story tellings.

To start with, I was interested to know the number of vehicles per household which involves the attribute **VOIT** which represents the number of cars in the household
Selecting the longest frequent itemset which contain the VOIT attribute
[(ANARR, Z), (DEROU, Z), (CATL, 1), (IMMI, 2.0), (INAT, 11.0), (NATC, 0.0), **(VOIT, 1.0),** (PNAI12, 1.0)]
and comparing it to also the longest itemset which contains also the attribute VOIT but with different value
[(ANARR, Z), (DEROU, Z), (CATL, 1), (IMMI, 2.0), (INAT, 11.0), **(INFAM, 1.0)**, (NATC, 0.0), **(VOIT, 2.0),** (PNAI12, 1.0)]

This shows that the only difference between these 2 itemsets is the attribute INFAM and it indicates that most households which contains 2 cars are consist of one family while as the household which have one car its not defined it may be single person or couple or maybe more than 1 family but anyways it is not stated because its not frequent there is a lot of possibilities for this.



itemset	length
[(ANARR, Z), (DEROU, Z), (CATL, 1), (IMMI, 2.0), (INAT, 11.0), (NATC, 0.0), (VOIT, 1.0), (PNAI12, 1.0)]	8.0
[(ANARR, Z), (DEROU, Z), (CATL, 1), (IMMI, 2.0), (INAT, 11.0), (INFAM, 1.0), (NATC, 0.0), (VOIT, 2.0), (PNAI12, 1.0)]	9.0

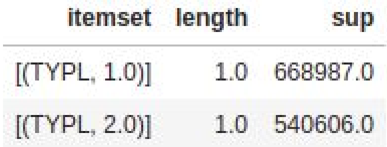
Secondly, searching for the itemsets which contains the attributes EMPL and ETUD gave 2 itemsets which contain only these 2 attributes but i chose to select the one having the emlp value as 16 not ZZ for having more meaningful analysis.
so from this item set **[(EMPL, 16), (ETUD, 2.0)]** we can derive that most people who have jobs with no time limit, permanent contract (CDI), civil servants are not registered in an educational institution. Moreover, comparing it to other itemsets which also contains the same values for EMPL and ETUD values shows that most people who possess these characteristics live in a main residence and also employed people which makes sense because they have jobs, this is derived from the itemset **[(CATL, 1), (EMPL, 16), (ETUD, 2.0), (TACT, 11.0)]** and other ones too you can check the jupyter notebook for more details.

Thirdly, I am interested in searching for everything related to country of birth, nationality and stuff of the people here in GRANEST i.e which are stated in the attributes **IMMI** which states the immigration situation, **INAT** whis is the Nationality indicator, **NATC** which shows the current condensed nationality and **PNAI12** which shows the country of birth in 12 positions. So first I selected the items which contain **PNAI12** attribute **[(PNAI12, 1.0)]** and it shows that most people are born in Metropolitan France and DOM-TOM-COM. Moreover, if you look at the other itemsets which contain this attribute you can see that the attributes **IMMI** of the value 2 and the attribute **NATC** of value 0 and also the attribute **INAT** of value 11 are all frequent which respectively means that most people who live in GrandEst are not immigrants, are french by birth.
For example :

[(ANARR, Z), (DEROU, Z), (CATL, 1), (EMPL, ZZ), (ILTUU, Z), **(IMMI, 2.0), (INAT, 11.0),** (INFAM, 1.0), **(NATC, 0.0)**, (TRANS, Z), **(PNAI12, 1.0)**]

[(ANARR, Z), (DEROU, Z), (CATL, 1), (EMPL, ZZ), (GARL, 1.0), (ILTUU, Z), (**IMMI, 2.0), (INAT, 11.0)**, **(NATC, 0.0)**, (TRANS, Z), (**PNAI12, 1.0)**]

Fourthly, in searching for the type of accommodation that most of the people in GrandEst have, i got 2 frequent itemsets that contain the attribute **TYPL** but one of them is more frequent than the other according to their support number one with value 1 which is a house is more frequent than from the one of value 2 which is an apartment. This indicated that the most frequent types of accommodation are a house and an appartment but the people who live in a house are more frequent than the people who live in an apartment.



itemset	length	sup
[(TYPL, 1.0)]	1.0	668987.0
[(TYPL, 2.0)]	1.0	540606.0

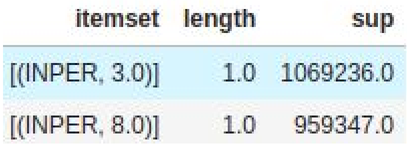
Digging deeper in the frequent itemsets by looking to **INFAM** attribute which as said before, it states the number of families in a household. It seems that one family in GrandEst can live in either an apartment or a house.

[(INFAM, 1.0), (TYPL, 1.0)]
[(INFAM, 1.0), (TYPL, 2.0)]

Fifthly, looking for the main mode of transport most often used for going to work using **TRANS** attribute shows that most people in GrandEst often go to work by Car, truck or even by van or other transport. And by taking the longest itemset which contain the attribute **TRANS** of value equal to 4 indicates that those people who uses oftenly these means of transport for going to work are indeed employed people either in real job or in apprenticeship or paid internship and mostly from their main residence.

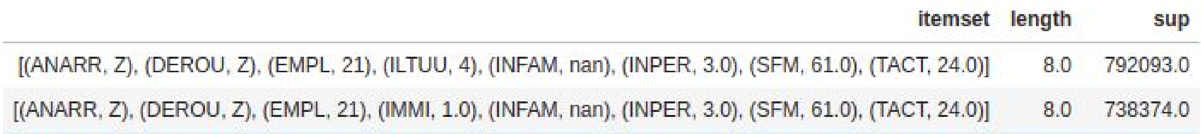
[(DEROU, Z), (CATL, 1), (RECH, Z), **(TACT, 11.0), (TRANS, 4)**]

Six, To study the overcrowded household property I had to look for the itemsets which contain the attribute **INPER** i.e the Number of persons in the household.
And there are 2 frequent values for this attribute which are 3 persons per household and 8 persons per household but the one with 3 is more frequent than the other.



itemset	length	sup
[(INPER, 3.0)]	1.0	1069236.0
[(INPER, 8.0)]	1.0	959347.0

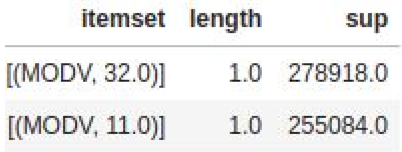
Then I wanted to study the characteristics for each, so starting for the most frequent one, the one with 3 persons I selected the longest itemsets which contain the attribute **INPER** with value 3.



itemset	length	sup
[(ANARR, Z), (DEROU, Z), (EMPL, 21), (ILTUU, 4), (INFAM, nan), (INPER, 3.0), (SFM, 61.0), (TACT, 24.0)]	8.0	792093.0
[(ANARR, Z), (DEROU, Z), (EMPL, 21), (IMMI, 1.0), (INFAM, nan), (INPER, 3.0), (SFM, 61.0), (TACT, 24.0)]	8.0	738374.0

And it seems also that the household which consist of 3 have the attribute **SFM** which describes the structure of the family equal to 61 i.e the household consist of two families with or without single person (s): two couples with or without children. Further they share also the attribute **TACT** with value of 24 i.e the household consists of housewives or men, in addition to the attribute **EMPL** with values 21 i.e they are self-employed. And from the second itemsets the **IMMI** attribute states that most of them are immigrants. I can guess that the 3 persons may not be related and they share the same flat even if it is not shown.

Seven, Studying the mode of live of the people in GrandEst by searching for MODC attribute, it seems that the most frequent way of life of the people in GrandEst is either Children of a couple (11) or Members of a couple with children (32).
the second is more frequent than the first.



itemset	length	sup
[(MODV, 32.0)]	1.0	278918.0
[(MODV, 11.0)]	1.0	255084.0

Taking the longest itemset with **MODV** value of 32

[(DEROU, Z), (**ETUD, 2.0), (INFAM, 1.0**), (**MODV, 32.0)]**

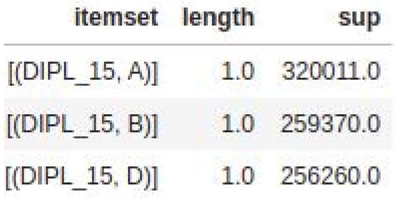
and the longest itemset with **MODV** value of 11

[(DEROU, Z), **(INFAM, 1.0), (MODV, 11.0)**]

[(DEROU, Z), **(IMMI, 2.0), (MODV, 11.0)**]

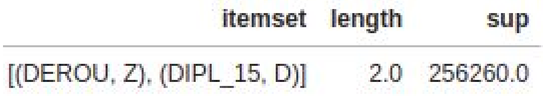
shows that most of the parents in GrandEst are not registered in an educational institution and that they live as a family (**INFAM** = 1). Moreover, the children live with their family and they are not immigrants.

Eight, What is the Highest diploma of the people living in GrandEst? From selecting the itemset which contain the attribute (**DIPL_15**).

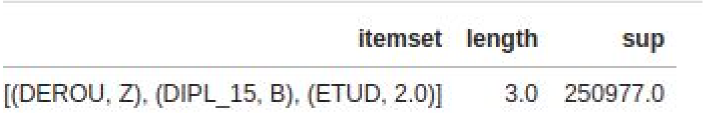


itemset	length	sup
[(DIPL_15, A)]	1.0	320011.0
[(DIPL_15, B)]	1.0	259370.0
[(DIPL_15, D)]	1.0	256260.0

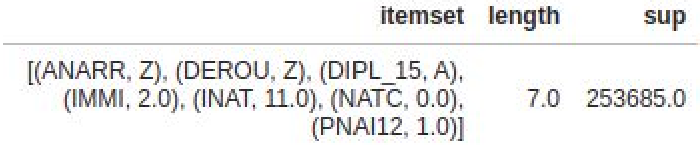
There are 3 frequent types of highest diploma that people in GrandEst possess.
A No diploma or at best BEPC, college certificate or DNB which is the most frequent.
B CAP, BEP C Baccalaureate (general, technological, professional) which is less frequent than A.
D Graduate Diploma which is slightly less frequent than B.
Taking the longest itemset which contain **DIPL_15** attribute for each type.



itemset	length	sup
[(DEROU, Z), (DIPL_15, D)]	2.0	256260.0



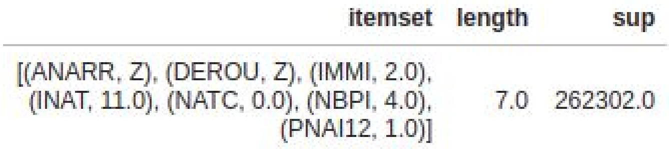
itemset	length	sup
[(DEROU, Z), (DIPL_15, B), (ETUD, 2.0)]	3.0	250977.0



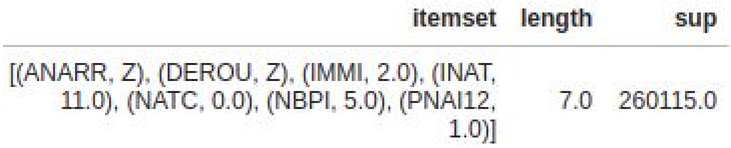
itemset	length	sup
[(ANARR, Z), (DEROU, Z), (DIPL_15, A), (IMMI, 2.0), (INAT, 11.0), (NATC, 0.0), (PNAI12, 1.0)]	7.0	253685.0

The results that I got shows that the people who have no diploma are not immigrants they are totally french and the ones who have baccalaureate are not registered in educational institute (second pic) probably they got it and started working .

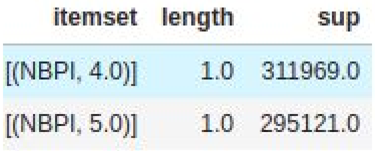
Nine, studying the number of rooms in the accommodation (NBPI).



itemset	length	sup
[(ANARR, Z), (DEROU, Z), (IMMI, 2.0), (INAT, 11.0), (NATC, 0.0), (NBPI, 4.0), (PNAI12, 1.0)]	7.0	262302.0



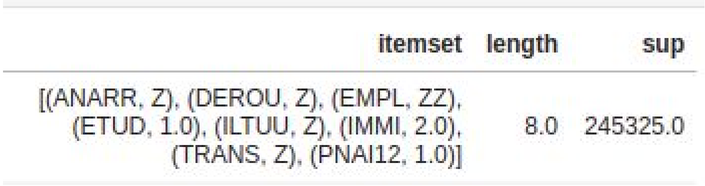
itemset	length	sup
[(ANARR, Z), (DEROU, Z), (IMMI, 2.0), (INAT, 11.0), (NATC, 0.0), (NBPI, 5.0), (PNAI12, 1.0)]	7.0	260115.0



itemset	length	sup
[(NBPI, 4.0)]	1.0	311969.0
[(NBPI, 5.0)]	1.0	295121.0

Most people's accommodations consists of 4 and 5 rooms in GrandEst and they are totally french people.

Finally, Most people in GrandEst who are registered in an educational institute are french and they live with their families i.e probably students .



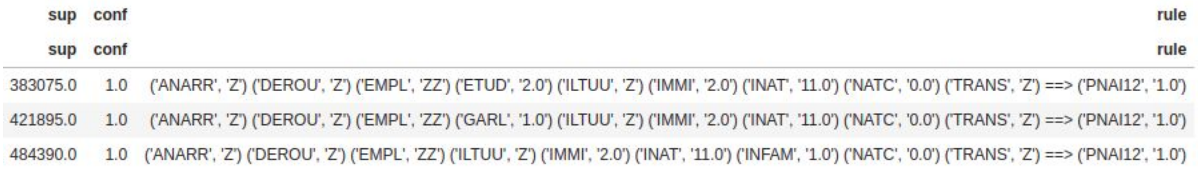
itemset	length	sup
[(ANARR, Z), (DEROU, Z), (EMPL, ZZ), (ETUD, 1.0), (ILTUU, Z), (IMMI, 2.0), (TRANS, Z), (PNAI12, 1.0)]	8.0	245325.0

Association Rules Analysis

To derive characteristics of the total population association analysis must be done. **Association analysis** is useful for discovering interesting relationships hidden in large **data** sets. The uncovered relationships can be represented in the form of **association** rules or sets of frequent items.

So first, I was interested to find characteristics of the people in GrandEst who are born in France métropolitaine et DOM-TOM-COM.

Selecting the association rules which have only the attribute PNAI12 in the right side of the rule at the same time have the maximum number of attributes to the left side of the rule and also having confidence of 100%, in this way I am able to collect the maximum number of characteristics of the people in GrandEst who are born in France métropolitaine et DOM-TOM-COM . As you can see there are a lot in common between the rules.



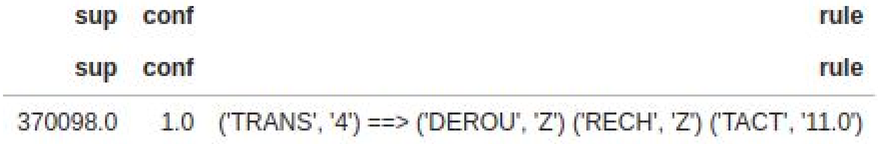
sup	conf	rule
sup	conf	rule
383075.0	1.0	('ANARR', 'Z') ('DEROU', 'Z') ('EMPL', 'ZZ') ('ETUD', '2.0') ('ILTUU', 'Z') ('IMMI', '2.0') ('INAT', '11.0') ('NATC', '0.0') ('TRANS', 'Z') ==> ('PNAI12', '1.0')
421895.0	1.0	('ANARR', 'Z') ('DEROU', 'Z') ('EMPL', 'ZZ') ('GARL', '1.0') ('ILTUU', 'Z') ('IMMI', '2.0') ('INAT', '11.0') ('NATC', '0.0') ('TRANS', 'Z') ==> ('PNAI12', '1.0')
484390.0	1.0	('ANARR', 'Z') ('DEROU', 'Z') ('EMPL', 'ZZ') ('ILTUU', 'Z') ('IMMI', '2.0') ('INAT', '11.0') ('INFAM', '1.0') ('NATC', '0.0') ('TRANS', 'Z') ==> ('PNAI12', '1.0')

since there are 3 different rules its best to select the one with highest support

('ANARR', 'Z') ('DEROU', 'Z') ('EMPL', 'ZZ') ('ILTUU', 'Z') (**'IMMI', '2.0') ('INAT', '11.0') ('INFAM', '1.0') ('NATC', '0.0')** ('TRANS', 'Z') ==> ('PNAI12', '1.0')

we get not very useful information from this, its just that people who are french by birth and not immigrants are born in France metropolitaine, which is something logical.And also with INFAM=1 states that there is one family in the household of those people.
but taking the first one with less support indicates that those who are not registered in an educational institute are born in France metropolitaine.

Secondly, I am interested to find the characteristics of the people who uses the car i.e TRANS=4 as a Main mode of transport most often used for going to work



sup	conf	rule
sup	conf	rule
370098.0	1.0	('TRANS', '4') ==> ('DEROU', 'Z') ('RECH', 'Z') ('TACT', '11.0')

so , the people who use the car for going to work are employed people either in real job or internship..

Thirdly, TYPMR attribute study the type of household of people.

('ANARR', 'Z') **('INAT', '11.0') ('TYPMR', '41.0')** ==> ('DEROU', 'Z') **('IMMI', '2.0') ('NATC', '0.0') ('PNAI12', '1.0')**

This shows that non immigrant families, but totally french ones are composed of a couple where only one man has the status of 'active with employment'

Fourth, VOIT attribute,

('ANARR', 'Z') ('INAT', '11.0') ('INFAM', '1.0') ('VOIT', '2.0') ==> ('DEROU', 'Z') ('IMMI', '2.0') ('NATC', '0.0') ('PNAI12', '1.0')

as usual all about french people, so french people have 2 cars per household.

Fifth, INFAM attribute

('ANARR', 'Z') ('EMPL', 'ZZ') **('INAT', '11.0')** (**'INFAM', '1.0')** ==> ('DEROU', 'Z') ('ILTUU', 'Z') (**'IMMI', '2.0') ('NATC', '0.0')** ('TRANS', 'Z') **('PNAI12', '1.0')**

most households in GrandEst consist of one family and they are french.

Six, GARL and VOIT attributes together

('GARL', '1.0') ('VOIT', '2.0') ==> ('DEROU', 'Z')

This one is interesting, people who own cars have parking space would most certainly have also a deux-roues à moteur du ménage.

Seven, ETUD and TACT attributes

('EMPL', '16') ('ETUD', '2.0') ==> ('TACT', '11.0')

people who have Jobs with no time limit, permanent contract,or civil servants and are not registered in an educational institute are Employed people, including apprenticeship or paid internship.

I can Explain all the rules but with the results that I have they will all be repetition of each other all the important and unique information are provided and no more than that can be derived from the results.