Valentin Larrieu

# Data analysis – Smokers dataset

The goal of this work is to find interesting data/patterns from a raw dataset. The excel file given are the response of 258 persons to 24 questions. Some question have simple answers, some contains a list of answers. The data are not normalized, so the first step of our study is to filter the data

## Data Cleaning:

- Some of values where missing, so we configured our import of the excel file to fill blank space with NA. Then we filled those NA with the most probable value (the value the most represented in the column)
- Weight cleaning: some weight were containing letters (kg) so we used a regex to remove them. We splited them in 3 categories :
    - 1= 60- (light)
    - 2= 60-90 (medium)
    - 3= 90+ (heavy)
- Height cleaning: some height contained letters and some were really low (1.6) so we did the same manipulation as for the weight, and scale up the really low values. We splited them in 3 categories :
    - 1= 140 (short)
    - 2= 140-180 (medium)
    - 3= 180+ (tall)
- Age Cleaning: we do 3 categories of age:
    - 1= 30 (Young)
    - 2= 30-50 (medium)
    - 3= 50+ (Senior)
- First cigarette cleaning: 4 categories
    - 1= 15- (VeryYoung)
    - 2= 15-21 (Young)
    - 3= 22-30 (Adult)
    - 4= 30+ (later)
- Gender cleaning: Female/Male has been changed to 1 / 2
- Phone cleaning: some different values were changed to match the other (uppercase difference for instance)
- Education & Family cleaning: both were converted to numerical

```
[1] Undergraduate degree (BachelorŠ–Ès)
[2] Graduate degree (MasterŠ–Ès, PhD)
[3] High school and/or vocational training
```

- Health condition – Method to stop smoking – Reason to quit smoking: Those column where containing list, separated with coma and with false values. So we cleared the false values and

separated each of the feature in a new column populated with corresponding TRUE/FALSE statement

- Reduce or stop smoking : String cleaning of truncated ones
- Contry: Contry grouped in 3 categories
    - 1= Other
    - 2= UK
    - 3= US
- Cigarette Everyday: we decided to split the set in 4 categories
    - 1= I do not smoke everyday
    - 2= 10 or fewer
    - 3= 20-30
    - 4= 31 or more
- After wake up cigarette: We spited the set in 4 categories
    - 1= within 5 minutes
    - 2= 5-60min
    - 3= 60min or more
    - 4= 2h or more
- Last time stop smoking: 4 categories
    - 1= 1 month
    - 2= 1 year
    - 3= 5 years ago
    - 4= Never tried
- Stop using the method listed: 3 categories
    - 1= No
    - 2= Short break
    - 3= Reduce
- Friends & Family: 5 categories
    - 1= Non-smokers
    - 2= Social smokers
    - 3= Moderate
    - 4= Heavy smokers
    - 5= Other
- Brand of cigarettes: just converted to numerical:

```
 [1] "Marlboro"     "Allure"       "Dunhill"      "Davidoff"
 [5] "Winston"      "Gitane"       "Kent"         "Parliament"
 [9] "Cedars"       "Gauloises"    "Camel"        "Winchester"
[13] "Other"        "Superkings"   "Lucky Strike" "Rothmans"
```

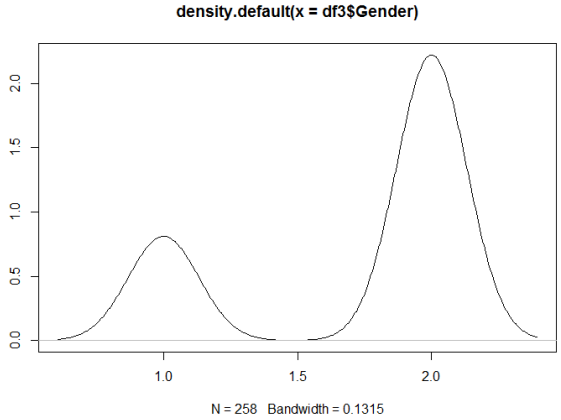
- Type of cigarette Box:
    - 1: Roll up
    - 2: 20 per pack (and the one which said 20)
- Own lighter:
    - 1: Not important
    - 2: Vital
- Salary:
    - 1: Not told
    - 2: 1000-
    - 3: 1000-5000
    - 4: 5000-10000

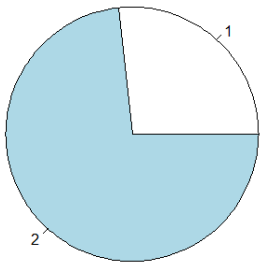
    - 5: 10000+
- Health:
    - 1: Healthy
    - 2: Major problem
    - 3: Minor problem
- Wanting to reduce or stop:
    - 1: No
    - 2: Reduce
    - 3: Stop
- Finally we convert all the table to numerical and rename the column

# Data Analysis:

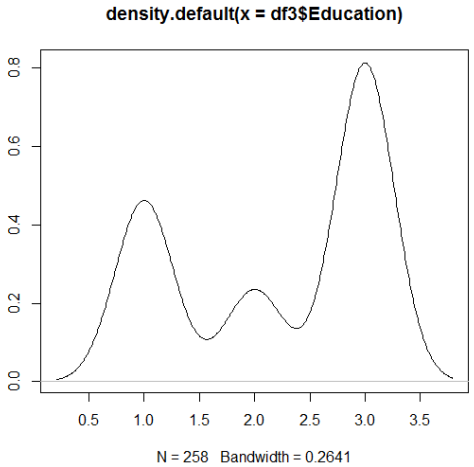
Some statistics:

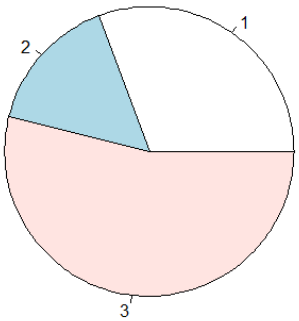
- Gender
    - As we can see a majority of the people of the dataset are male (label 2: 73%)



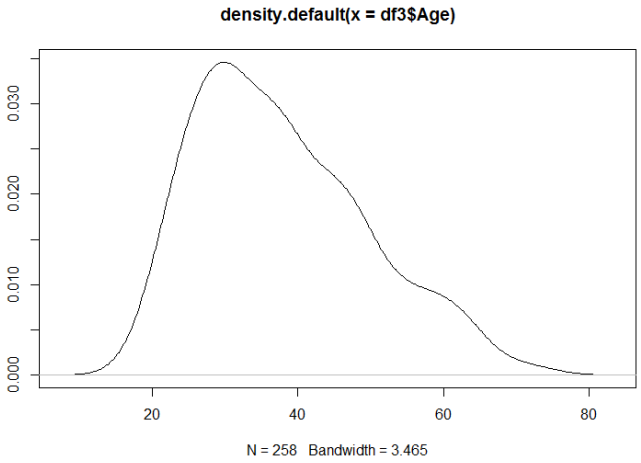




- Education
    - As we can see there is more undergratuated(label 3 : 53.3%), then comes the one graduated (label 1 : 32.6%) and finaly High school and vocational training(15.1%)



- Age
    - We can see that the people answering the questions are for a lot of them between 20 and 40



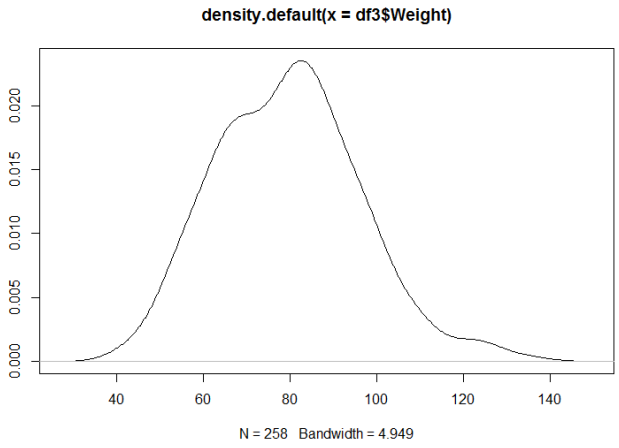

    - The repartition is :

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
18.00   29.00   36.50   37.91   46.00   73.00
```

    - Std

```
             Age
11.68858634
```

- Weight
    - A big part of the set is between 60 and 90 kg



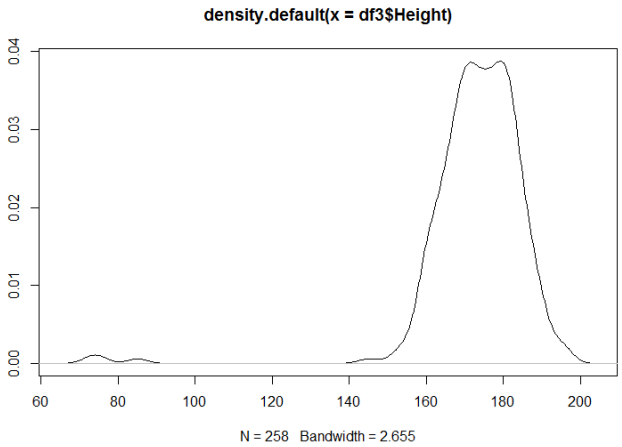

    - Repartition

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
42.00   67.00   80.00   79.63   90.00  135.00
```

    - Std

```
      Weight
16.69455739
```

- Height
    - The set is for the most between 160 and 190cm



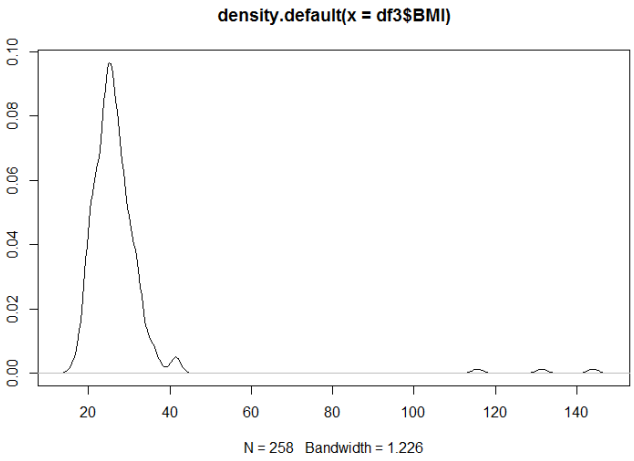

    - Repartition

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
73.0   168.0   174.0   173.1   180.0   196.0
```

    - Std

```
      Height
13.56209640
```

- BMI
    - BMI for most between 25 and 35



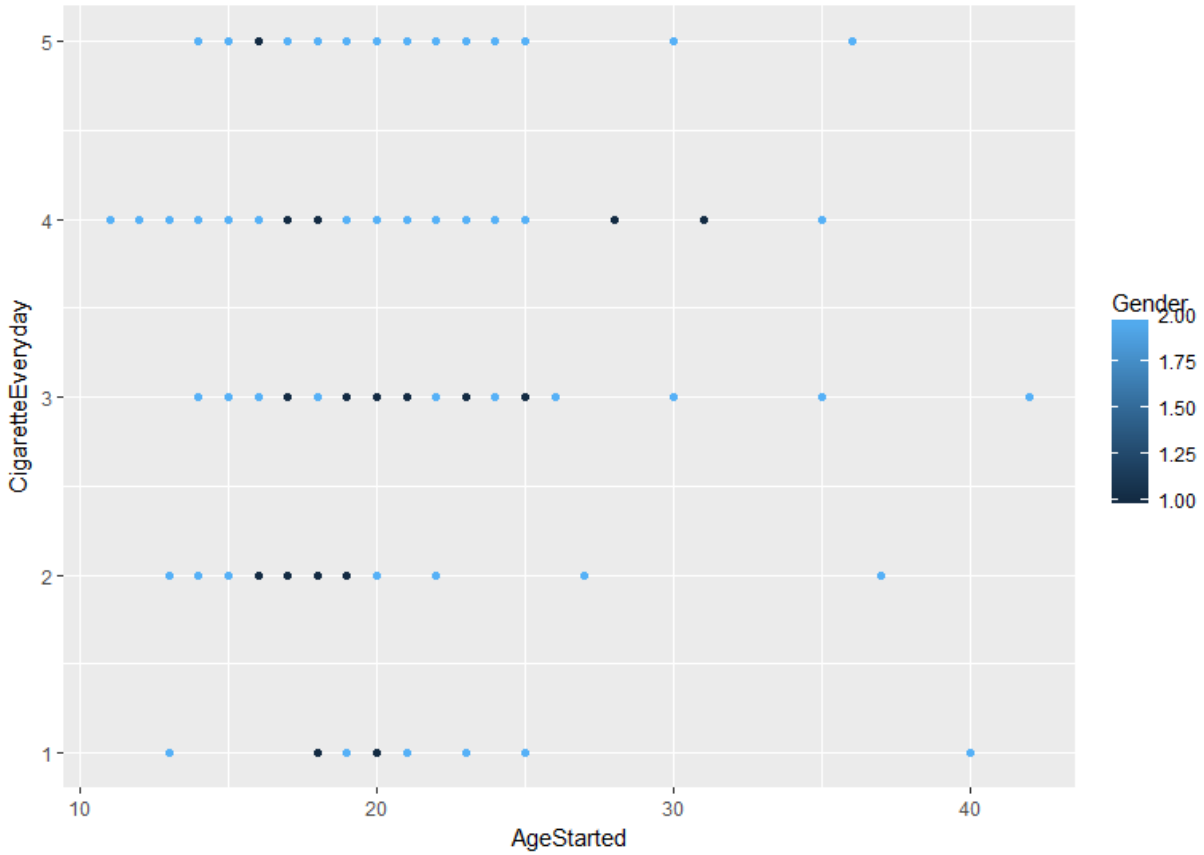

    - Repartition

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
16.73   23.18   25.67   27.32   28.73  143.90
```

    - Std

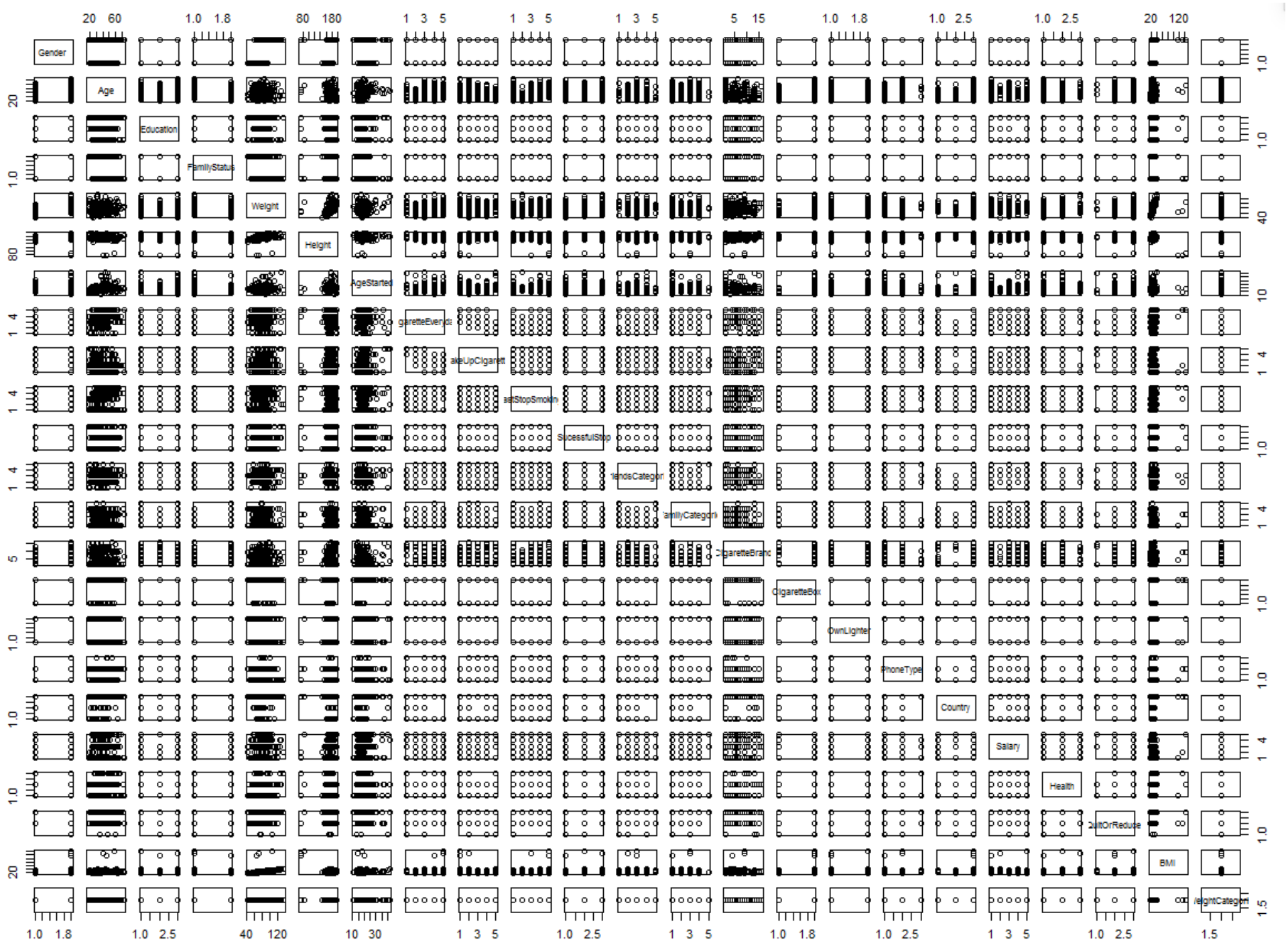
```
             BMI
12.12526003
```

Some link between data:




- We can see that men seem to tend to start earlier than woman (for this set at least)
- We can also see that the people starting to smoke the earlier tend to smoke a lot alter (here the 4rth categorie is between 21 and 31 per day)

We also did a pair of most our data to visualise eventual clusters:


Gender
Age
Education
FamilyStatus
Weight
Height
AgeStarted
SucessfulStop
CigaretteBrand
CigaretteBox
OwnLighter
PhoneType
Country
Salary
Health
BMI

Correlation matrix:

| | Gender | Age | Education | FamilyStatus | Weight | Height | AgeStarted | CigaretteEveryday | wakeUpCigarette |
|---|---|---|---|---|---|---|---|---|---|
| Gender | 1.000000000 | -0.010421536 | -0.137395533 | -0.126671680 | 0.616254694 | 0.375230141 | 0.056472358 | 0.07809657 | 0.013597573 |
| Age | -0.010421536 | 1.000000000 | -0.032086158 | -0.396658656 | 0.013058836 | -0.181018787 | 0.234995037 | 0.23779086 | -0.225009644 |
| Education | -0.137395533 | -0.032086158 | 1.000000000 | -0.115348623 | -0.138086539 | -0.102227996 | -0.078624395 | 0.08518079 | -0.071581659 |
| FamilyStatus | -0.126671680 | -0.396658656 | -0.115348623 | 1.000000000 | -0.212683177 | 0.011134677 | -0.259008229 | -0.21339020 | 0.083594499 |
| Weight | 0.616254694 | 0.013058836 | -0.138086539 | -0.212683177 | 1.000000000 | 0.383093591 | 0.180463079 | 0.06878954 | 0.096239044 |
| Height | 0.375230141 | -0.181018787 | -0.102227996 | 0.011134677 | 0.383093591 | 1.000000000 | 0.054707257 | -0.14091483 | 0.167408919 |
| AgeStarted | 0.056472358 | 0.234995037 | -0.078624395 | -0.259008229 | 0.180463079 | 0.054707257 | 1.000000000 | -0.03383219 | 0.189895771 |
| CigaretteEveryday | 0.078096565 | 0.237790858 | 0.085180794 | -0.213390195 | 0.068789543 | -0.140914835 | -0.033832192 | 1.00000000 | -0.585234239 |
| wakeUpCigarette | 0.013597573 | -0.225009644 | -0.071581659 | 0.083594499 | 0.096239044 | 0.167408919 | 0.189895771 | -0.58523424 | 1.000000000 |
| LastStopSmoking | 0.106820349 | 0.119222665 | -0.014560308 | -0.002573211 | 0.093849033 | -0.064611285 | 0.008914732 | 0.09950337 | -0.052669125 |
| SucessfulStop | 0.043302951 | 0.081557544 | -0.030253879 | -0.058574085 | 0.076650415 | 0.039024132 | 0.029153764 | -0.08544399 | 0.070216195 |
| FriendsCategorie | -0.055207665 | -0.027874661 | 0.002263609 | 0.034372126 | 0.007793753 | 0.009629693 | 0.003208239 | 0.09488736 | -0.026103962 |
| FamilyCategorie | -0.101231675 | -0.050500062 | 0.029454288 | 0.030673271 | -0.112426217 | -0.127255625 | -0.059743696 | 0.10053453 | -0.090612958 |
| CigaretteBrand | 0.056812534 | -0.002123576 | 0.015784467 | 0.088604366 | 0.077674534 | 0.059557404 | -0.025201838 | -0.07498734 | 0.039372843 |
| CigaretteBox | 0.016006539 | -0.026047094 | -0.053519879 | 0.048929606 | -0.009619880 | -0.023097549 | 0.096007783 | 0.11219183 | -0.001969798 |
| OwnLighter | -0.052023452 | -0.050077834 | -0.092430321 | 0.048807422 | -0.043722477 | -0.016964400 | 0.053619186 | -0.07692123 | 0.065339411 |
| PhoneType | 0.004595777 | -0.084269545 | 0.040775189 | 0.073469433 | -0.073175047 | 0.024130295 | -0.103343353 | -0.10160984 | 0.099382981 |
| Country | -0.134996810 | 0.086173215 | 0.065085234 | -0.009215187 | -0.006057558 | -0.124487235 | 0.103797794 | 0.27357779 | -0.171693152 |
| Salary | 0.176371029 | 0.011468411 | -0.102071131 | -0.031742719 | 0.103015173 | 0.079352459 | 0.102349182 | -0.14134253 | 0.134801403 |
| Health | 0.010644713 | 0.233947280 | -0.103352469 | -0.049195560 | 0.146039975 | -0.128807768 | 0.045916003 | 0.16225952 | -0.178002922 |
| QuitOrReduce | 0.082134590 | 0.013307293 | -0.039213087 | -0.031153896 | 0.136732688 | 0.075518803 | -0.071897867 | 0.06574409 | -0.122344304 |
| BMI | 0.210321130 | 0.132640449 | -0.051796499 | -0.134198265 | 0.337343480 | -0.663974249 | 0.027780951 | 0.16439134 | -0.078073152 |

| | LastStopSmoking | SucessfulStop | FriendsCategorie | FamilyCategorie | CigaretteBrand | CigaretteBox | OwnLighter | PhoneType | Country |
|---|---|---|---|---|---|---|---|---|---|
| Gender | 0.106820349 | 0.043302951 | -0.0552076651 | -0.101231675 | 0.056812534 | 0.016006539 | -0.05202345 | 0.004595777 | -0.134996810 |
| Age | 0.119222665 | 0.081557544 | -0.0278746613 | -0.050500062 | -0.002123576 | -0.026047094 | -0.05007783 | -0.084269545 | 0.086173215 |
| Education | -0.014560308 | -0.030253879 | 0.0022636092 | 0.029454288 | 0.015784467 | -0.053519879 | -0.09243032 | 0.040775189 | 0.065085234 |
| FamilyStatus | -0.002573211 | -0.058574085 | 0.0343721260 | 0.030673271 | 0.088604366 | 0.048929606 | 0.04880742 | 0.073469433 | -0.009215187 |
| Weight | 0.093849033 | 0.076650415 | 0.0077937527 | -0.112426217 | 0.077674534 | -0.009619880 | -0.04372248 | -0.073175047 | -0.006057558 |
| Height | -0.064611285 | 0.039024132 | 0.0096296926 | -0.127255625 | 0.059557404 | -0.023097549 | -0.01696440 | 0.024130295 | -0.124487235 |
| AgeStarted | 0.008914732 | 0.029153764 | 0.0032082393 | -0.059743696 | -0.025201838 | 0.096007783 | 0.05361919 | -0.103343353 | 0.103797794 |
| CigaretteEveryday | 0.099503373 | -0.085443987 | 0.0948873622 | 0.100534527 | -0.074987342 | 0.112191833 | -0.07692123 | -0.101609842 | 0.273577792 |
| wakeUpCigarette | -0.052669125 | 0.070216195 | -0.0261039623 | -0.090612958 | 0.039372843 | -0.001969798 | 0.06533941 | 0.099382981 | -0.171693152 |
| LastStopSmoking | 1.000000000 | 0.097595141 | -0.0410780057 | -0.097537280 | 0.109164602 | -0.046714381 | -0.04885012 | -0.003400261 | 0.011634484 |
| SucessfulStop | 0.097595141 | 1.000000000 | -0.0844880393 | 0.007930185 | 0.141350172 | -0.035989462 | 0.01357047 | 0.012849170 | -0.011599960 |
| FriendsCategorie | -0.041078006 | -0.084488039 | 1.0000000000 | 0.151011932 | -0.152998310 | 0.130803066 | -0.02219564 | 0.016423427 | 0.108614191 |
| FamilyCategorie | -0.097537280 | 0.007930185 | 0.1510119322 | 1.000000000 | -0.008808040 | 0.041849608 | -0.05286074 | 0.013206687 | 0.115332636 |
| CigaretteBrand | 0.109164602 | 0.141350172 | -0.1529983098 | -0.008808040 | 1.000000000 | -0.274256338 | 0.05377259 | -0.084442764 | -0.114279073 |
| CigaretteBox | -0.046714381 | -0.035989462 | 0.1308030662 | 0.041849608 | -0.274256338 | 1.000000000 | -0.03438990 | -0.018418914 | 0.165983114 |
| OwnLighter | -0.048850125 | 0.013570465 | -0.0221956382 | -0.052860744 | 0.053772591 | -0.034389896 | 1.00000000 | -0.021783696 | -0.077373888 |
| PhoneType | -0.003400261 | 0.012849170 | 0.0164234270 | 0.013206687 | -0.084442764 | -0.018418914 | -0.02178370 | 1.000000000 | -0.021191292 |
| Country | 0.011634484 | -0.011599960 | 0.1086141911 | 0.115332636 | -0.114279073 | 0.165983114 | -0.07737389 | -0.021191292 | 1.000000000 |
| Salary | 0.077759224 | -0.025398468 | 0.0096026728 | 0.006607498 | 0.079497653 | -0.161050797 | 0.09068698 | 0.114083140 | -0.148848399 |
| Health | 0.082785529 | -0.007467349 | -0.0273102265 | -0.028689343 | 0.022426582 | -0.035742742 | -0.20051882 | -0.050307818 | 0.103744161 |
| QuitOrReduce | 0.201365681 | 0.084976270 | -0.0002523913 | -0.043458712 | 0.035913371 | -0.035726551 | -0.08606902 | -0.042828166 | 0.013368920 |
| BMI | 0.128911520 | 0.054028015 | -0.0226281357 | 0.002859108 | 0.057232447 | 0.020696253 | -0.01893745 | -0.034704933 | 0.064039386 |

| | Salary | Health | QuitOrReduce | BMI |
|---|---|---|---|---|
| Gender | 0.176371029 | 0.010644713 | 0.0821345902 | 0.210321130 |
| Age | 0.011468411 | 0.233947280 | 0.0133072928 | 0.132640449 |
| Education | -0.102071131 | -0.103352469 | -0.0392130875 | -0.051796499 |
| FamilyStatus | -0.031742719 | -0.049195560 | -0.0311538957 | -0.134198265 |
| Weight | 0.103015173 | 0.146039975 | 0.1367326882 | 0.337343480 |
| Height | 0.079352459 | -0.128807768 | 0.0755188029 | -0.663974249 |
| AgeStarted | 0.102349182 | 0.045916003 | -0.0718978667 | 0.027780951 |
| CigaretteEveryday | -0.141342528 | 0.162259516 | 0.0657440942 | 0.164391339 |
| wakeUpCigarette | 0.134801403 | -0.178002922 | -0.1223443044 | -0.078073152 |
| LastStopSmoking | 0.077759224 | 0.082785529 | 0.2013656815 | 0.128911520 |
| SucessfulStop | -0.025398468 | -0.007467349 | 0.0849762704 | 0.054028015 |
| FriendsCategorie | 0.009602673 | -0.027310227 | -0.0002523913 | -0.022628136 |
| FamilyCategorie | 0.006607498 | -0.028689343 | -0.0434587123 | 0.002859108 |
| CigaretteBrand | 0.079497653 | 0.022426582 | 0.0359133707 | 0.057232447 |
| CigaretteBox | -0.161050797 | -0.035742742 | -0.0357265508 | 0.020696253 |
| OwnLighter | 0.090686978 | -0.200518818 | -0.0860690206 | -0.018937455 |
| PhoneType | 0.114083140 | -0.050307818 | -0.0428281660 | -0.034704933 |
| Country | -0.148848399 | 0.103744161 | 0.0133689201 | 0.064039386 |
| Salary | 1.000000000 | -0.003994587 | -0.0713885585 | 0.040231003 |
| Health | -0.003994587 | 1.000000000 | 0.0552921869 | 0.181414857 |
| QuitOrReduce | -0.071388559 | 0.055292187 | 1.0000000000 | 0.006827916 |
| BMI | 0.040231003 | 0.181414857 | 0.0068279164 | 1.000000000 |

Link we could see with that matrix:

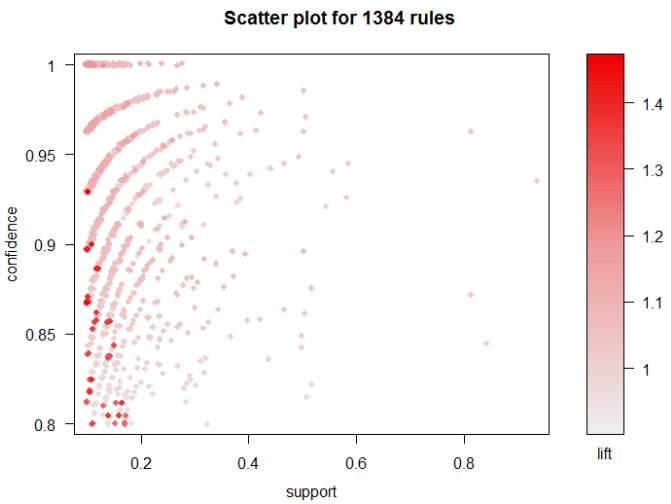
- Gender-Weight (0.62): As we could expect gender and weight are correlated
- BMI-Height (0.66): By contruction they are indeed linked, it is verified by the correlation (same as the weight)
- WakeUpCigarette-CigaretteEveryday (0.59): More interesting the moment you take your first cigarette seem to be correlated with the number of cigarette you consume everyday (the people starting to smoke earlier each day tend to smoke more cigarettes)

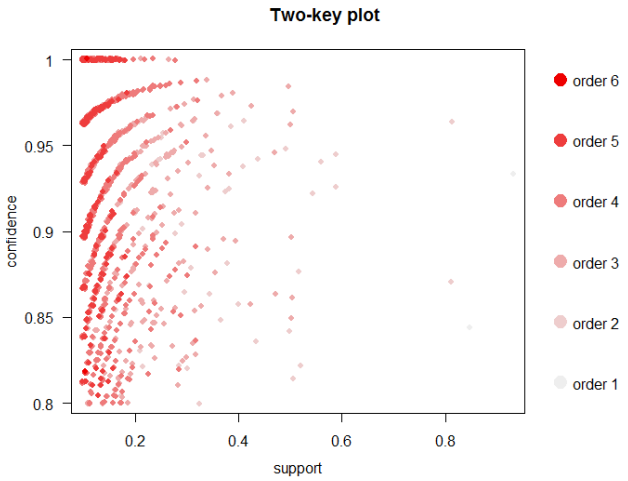
Then using the **apriori** algorithm:

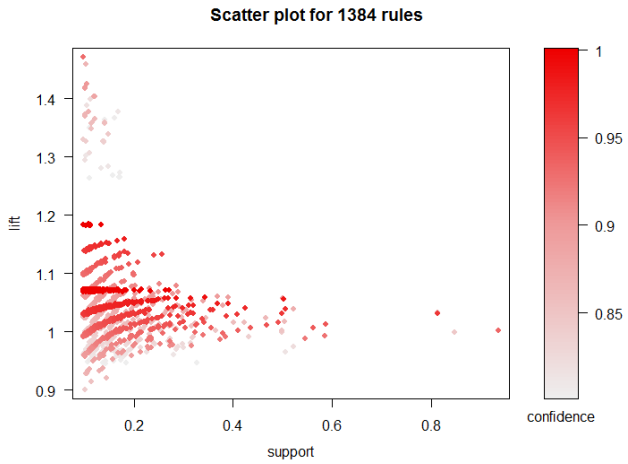
```
{wakeUpCigarette=4}                    => {Health=1}       0.101   0.812      1.302  26
```

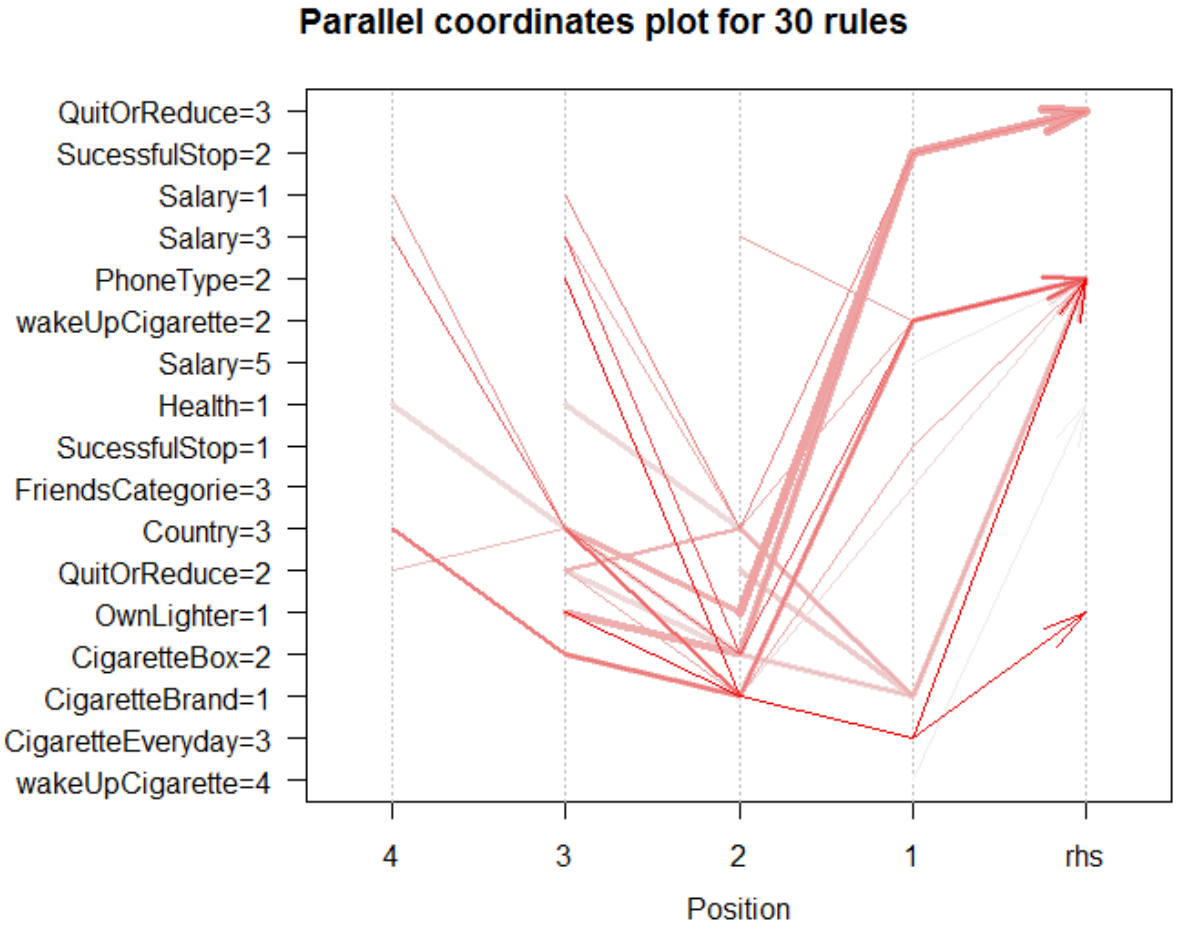
- The people taking they first cigarette of the day later tend to be the healthier

The rules generated look like this:





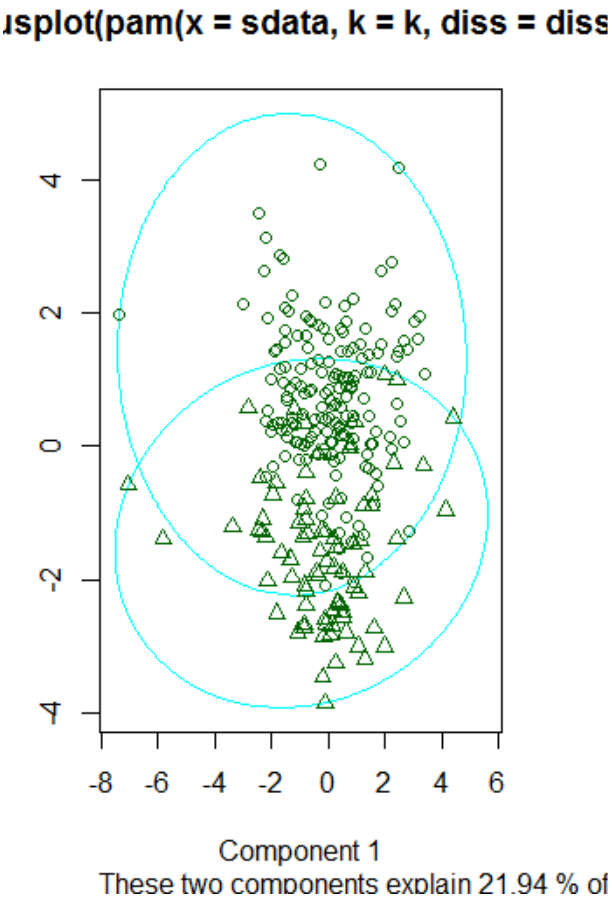



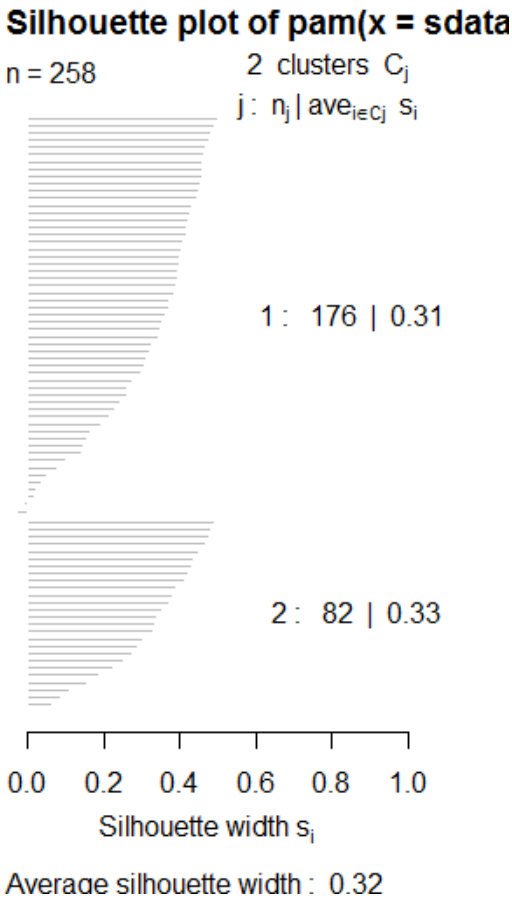




- Some logical result can be seen here: the people wanting the reduce or stop smoking ( QuitOrReduce=2 or 3) managed to stop smoking for a short period (SucessfulStop=2). It seems to show that motivation is important

We also tried other method:

- Clustering with medoids (here for BMI)

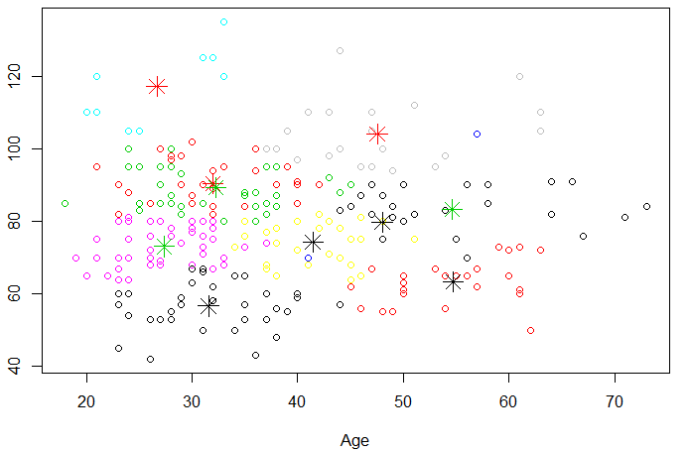



- Clustering with KMean :

```
within cluster sum of squares by cluster:
 [1] 4673.181 3541.857 5257.889 1582.024 1930.386 4757.932 3249.435 5799.661 5162.598 3485.070
 (between_SS / total_SS =  80.7 %)
```

(10 clusters)



- The result of those 2 methods where not really successful

As a conclusion, we can say that using R to see correlation in those data was interesting because it helped us to identify link we could not have seen with our bare eyes. Nonetheless, that study also showed us how hard and long the process of data cleaning is (it represented dozens of hour of work), and how experience, to know “where to find” can be useful.