Gastone Riccardo Tolli
ID: 33506151

Date: 29/08/2018

IS71069B Geometric Data Analysis

Assignment 1 (re-sit)

Introduction and Objectives

The dataset used is a music questionnaire found in the case studies section of the EnQuireR website (link: http://enquirer.free.fr/case-studies/Music/index.html). This questionnaire consists of 29 questions, therefore 29 variables, and 108 respondents.

The purpose of this paper is to make some exploratory analysis on the music listening behavior of the respondents based off the variables available. Observations on point of interest and conclusions on trends will be reported at the end of the analysis.

Data preparation and analysis

The software used for the analysis is R (version 3.5.0) used through RStudio.
The main library used is FactoMiner and factoextra which has been downloaded with the following command:

```
> install.packages("FactoMineR")
> install.packages("factoextra")
```

and loaded with:

```
> library(FactoMineR)
> library(factoextra)
```

Loading the dataset:

```
> data=read.csv("Music2.csv")
```

dimensions of dataframe:

```
> dim(data)
[1] 108  29
```

Gastone Riccardo Tolli
ID: 33506151

Date: 29/08/2018

The dataset contains 108 observations and 29 variables, each observation represent a person and the variables are:

```
> names(data)
 [1] "SEX"                                 "AGE"
 [3] "SPC"                                 "Kind.of.music"
 [5] "Why"                                 "Knowledge"
 [7] "Wished.budget"                       "Internet"
 [9] "Instrument"                          "Radio"
[11] "access"                              "Kind.of.shop"
[13] "Kind.of.material"                    "How.many.new.bands.by.month"
[15] "Frequency.listen.to.music.in.one.week" "New.bands.friends"
[17] "Website.discovering"                 "radio.discovering"
[19] "tv.show.discovering"                 "magazine.discovering"
[21] "parents.discovering"                 "poster.discovering"
[23] "movie.discovering"                   "concert.discovering"
[25] "enough.concert"                      "go.to.concert.of.known.bands
"

[27] "radio.influence"                     "same.music.as.friends"
[29] "listen.to.whole.album"
```

By looking at the data, some columns seem to have too little of a variation to make them statistically significant:

```
> attach(data)
> table(AGE)
AGE
17 18 19 20 21 22 23 25 37 43 48
 1 18 15 25 24 19  2  1  1  1  1
```

The bulk of the respondents go from 18 to 22 years old; in fact, they are 101/108 of the surveys. I am going to remove anyone that is not within this age range and then remove the column age altogether:

```
> data<-data[!(AGE<18 | AGE>22),]
> data<-data[,-2]
```

The column SPC looks as follows:

```
> table(SPC)
SPC
manager student
      0     101
```

Originally there were 2 managers which were removed as a consequence of the previous step (removing AGE outliers), 2 respondents is still extremely low, therefore I am removing this column too.

```
data<-data[,-2]
```

Gastone Riccardo Tolli
ID: 33506151

Date: 29/08/2018

Descriptive statistics

Here I will carry out a few descriptive statistics in order to give a general sense of the dataset.

The names of the columns are self-explanatory, therefore, there is no need for a variable description.

Distribution of male and female respondents:

```
> table(SEX)
SEX
 F  M 
76 25 
```

Distribution of the kind of music listened:

```
> table(Kind.of.music)
Kind.of.music
    blues    classic hard rock     other       pop       rap      rock 
        1          0         1        18        40         5        36 
```

A bar graph that shows the most and least popular music genres listened by the respondents.

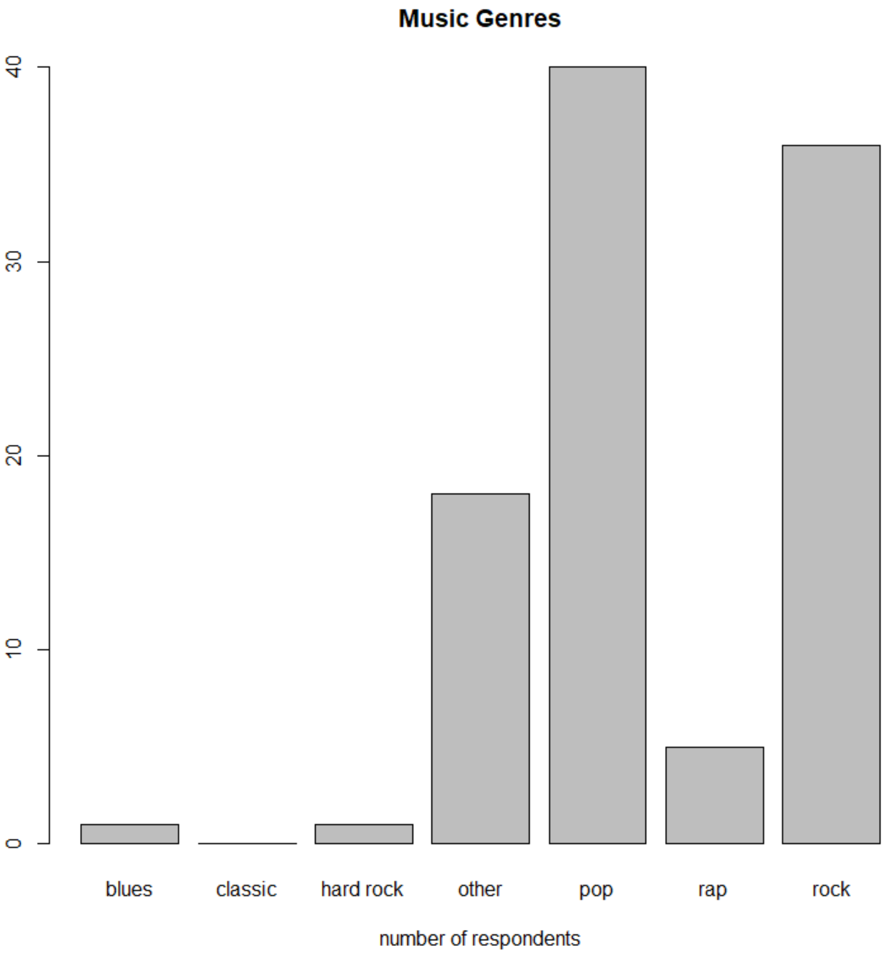


Fig.1

Gastone Riccardo Tolli
ID: 33506151

Date: 29/08/2018

In the column “Internet” it is possible to see that illegal downloading (53) is used more than twice as much than legal downloading (20).

```
> table(Internet)
Internet
                              e-shopping illegal downloading   legal downlo
ading
                 25                    3                  53
20
```

Multiple Correspondence Analysis

I am going to perform an MCA on a selection of variables. The goal is to find some trend revolving the illegal downloading.

Creating new dataset with the columns used for this analysis.

As during the analysis, the graphs were incredibly overcrowded I am removing some variables. Namely cntry from the contextual variables and all the variables after ipfrule.

```
> data.MCA=data[,c("SEX" ,   "Why"   ,   "Knowledge" , "Wished.budget" ,"Int
ernet", "Instrument")]
```

The criterion that was used to select these variables is to only take the ones that more likely contain information regarding the illegal downloading.

Carrying out the MCA

```
> MCA1=MCA(data.MCA, ncp = 5, ind.sup = NULL, quanti.sup = NULL ,quali.sup
= c(1,6), excl=NULL, graph = TRUE)
> var<-get_mca_var(MCA1)
```

Let us look at the explained variance and eigenvalues for the top 25 dimensions

```
> head(get_eig(MCA1), 100)
       eigenvalue variance.percent cumulative.variance.percent
Dim.1   0.3731948        12.439825                    12.43983
Dim.2   0.3454386        11.514619                    23.95444
Dim.3   0.3306855        11.022849                    34.97729
Dim.4   0.2963624         9.878747                    44.85604
Dim.5   0.2862436         9.541454                    54.39749
Dim.6   0.2712140         9.040466                    63.43796
Dim.7   0.2405042         8.016807                    71.45477
Dim.8   0.2084868         6.949559                    78.40433
Dim.9   0.1989185         6.630617                    85.03494
Dim.10  0.1821650         6.072166                    91.10711
Dim.11  0.1394890         4.649635                    95.75674
Dim.12  0.1272977         4.243255                   100.00000
```

The cumulative variance percent reaches 100 % at the 12th dimension. It seems like there are no overwhelmingly dominant axes. There is no single dimension that explain a lot of variance. The top 3 dimensions explain 35 % of the variance, and the top 5 dimensions 54.4 % while top 2 they explain 12.4 and 11.4 % respectively.

Now I will plot the explained variance

```
fviz_eig(MCA1, choice="variance", labels=TRUE)
```

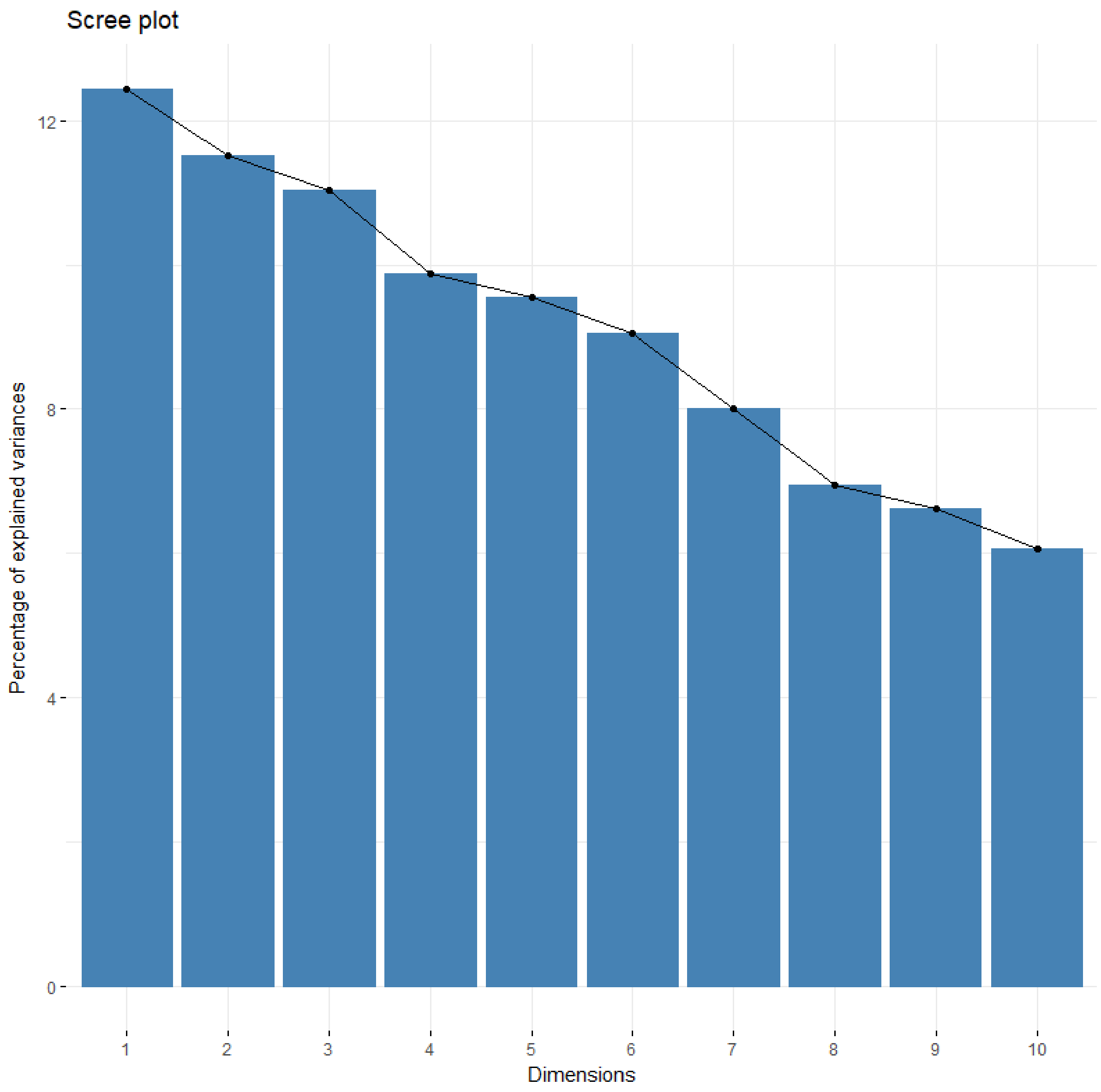
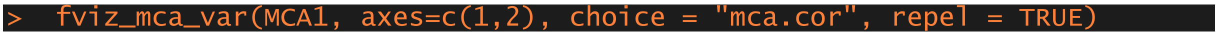


Fig. 2

The following charts is the squared correlation of coefficient of variables with regards to the first 2 dimensions.



```
>  fviz_mca_var(MCA1, axes=c(1,2), choice = "mca.cor", repel = TRUE)
```

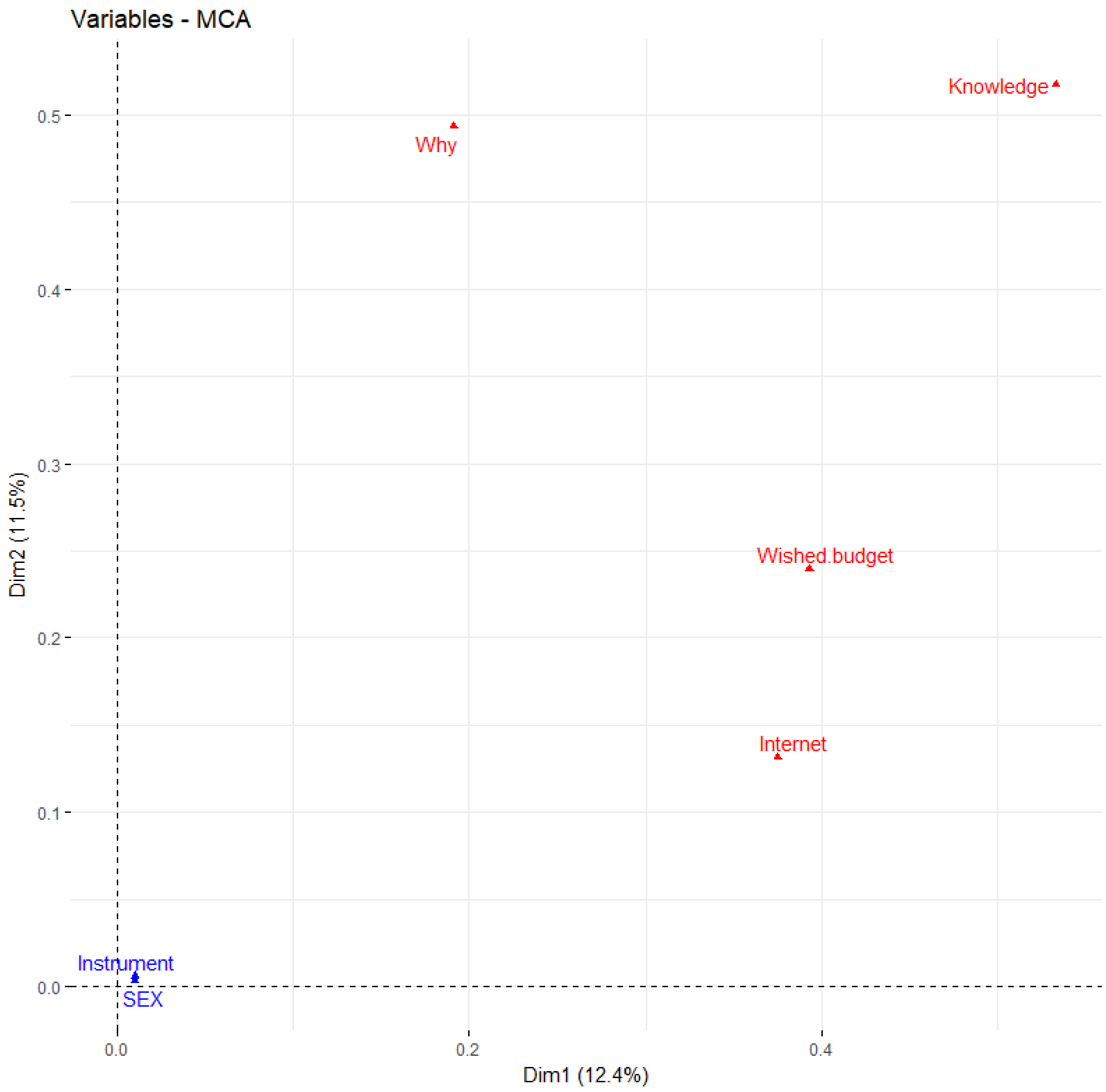
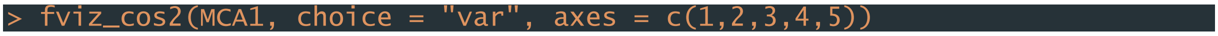


Fig.3

It is possible to see how there is a very low correlation between Instrument and SEX with the first two dimensions. While the other variables displayed in red have a higher correlation, Knowledge has it the most with both dimensions being in the top right corner.

The next visualization is the squared cosine. It shows the quality of representation of variable categories on the first 5 axes. The fifth axis was chosen as an arbitrary cut off point since the higher the axis the less variance it explains.



```
> fviz_cos2(MCA1, choice = "var", axes = c(1,2,3,4,5))
```

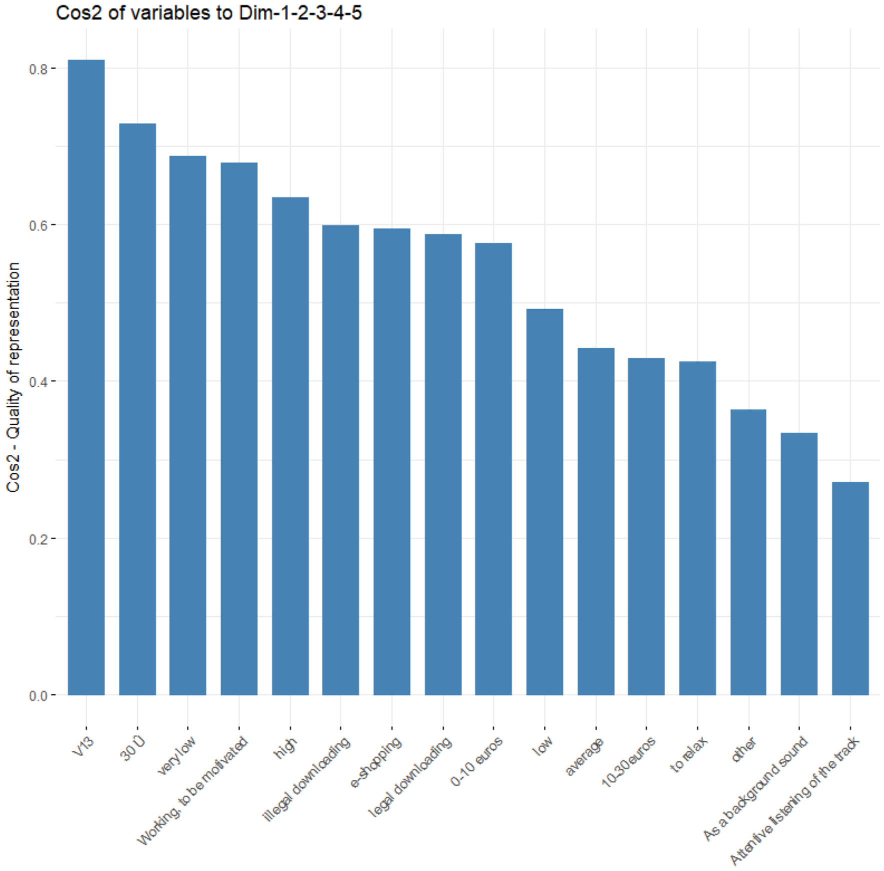


Fig.4

The coefficients for the variables seem to decrease fairly smoothly.

An alternative way of visualizing it:

```
> fviz_mca_var(MCA1, axes=c(1,2), col.var = "cos2", repel = TRUE)
```

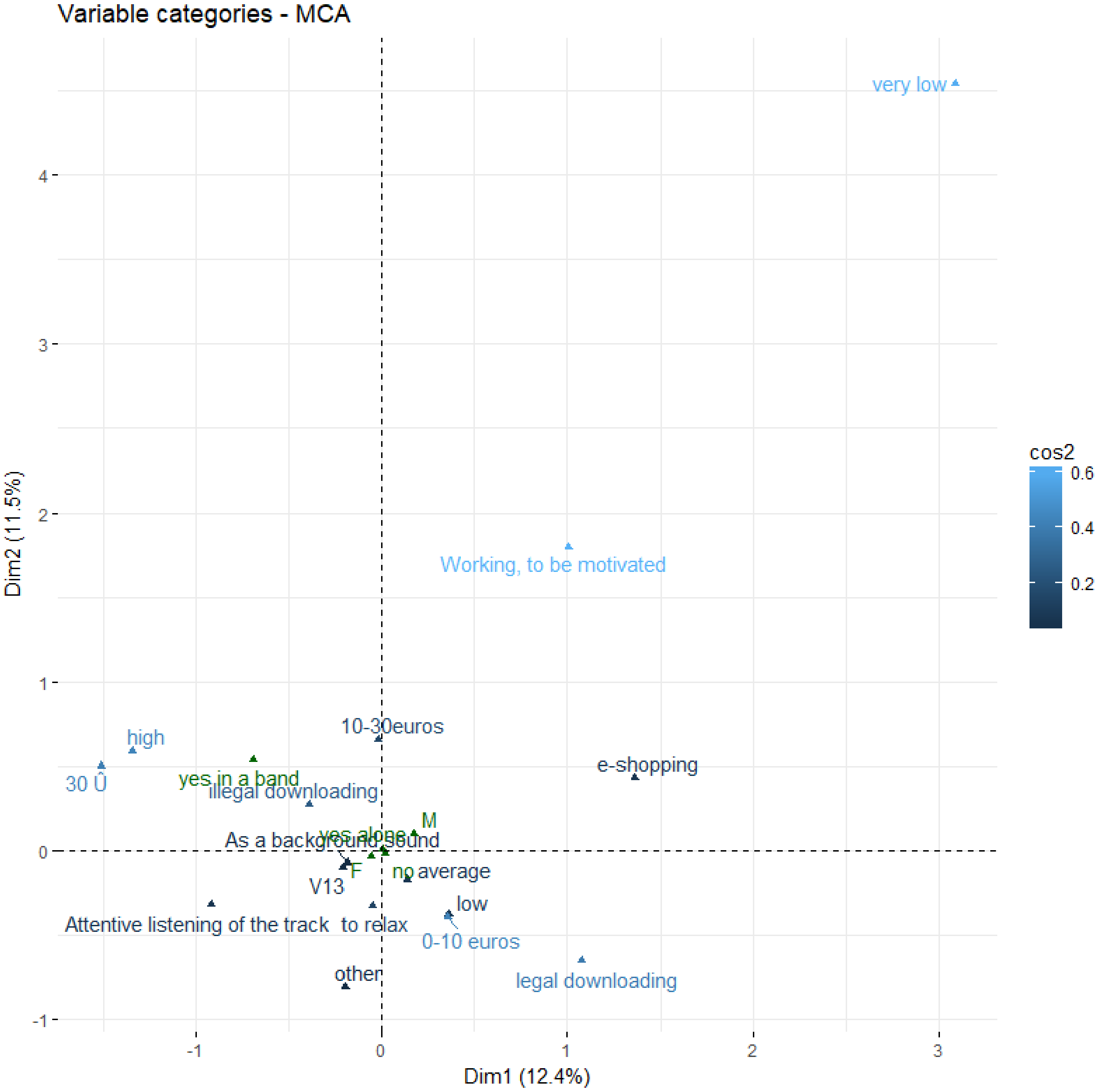


Fig.5

Now I will compute the point clouds of individuals.

```
> fviz_mca_ind(MCA1, col.ind = "cos2", axes=c(1,2), geom=c("point"), repel
=TRUE)
```

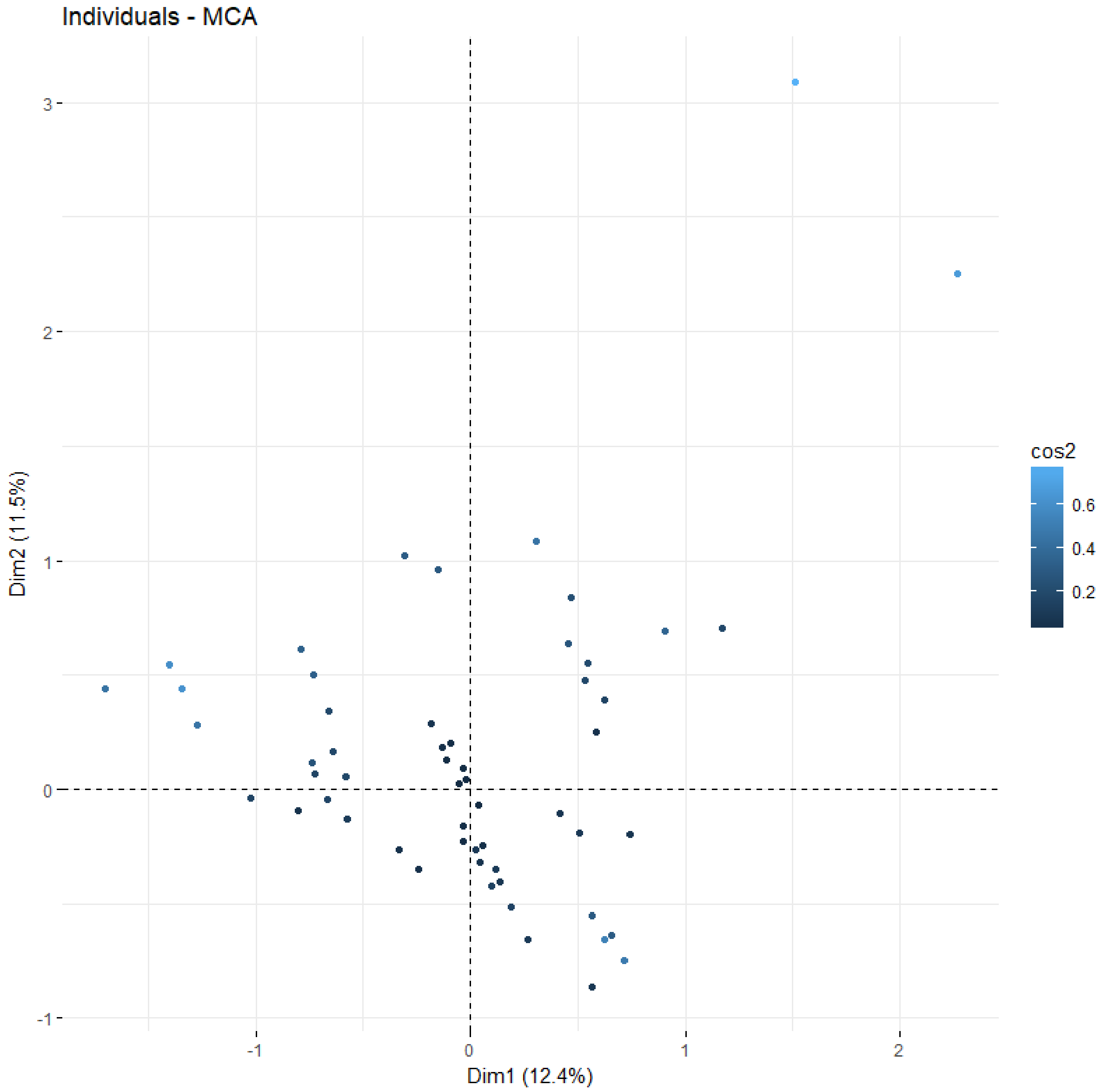


Fig.6

As we can see the points are relatively evenly spread but they gravitate towards the origin with the exception of very few outliers, they are fairly evenly distributed along both dimension 1 and 2.

Gastone Riccardo Tolli
ID: 33506151

Date: 29/08/2018

I am going to explore the coordinates of the variable categories on dimension 1 and 2.

```
> fviz_mca_var(MCA1, axes=c(1,2), geom=c("point", "text"), repel=TRUE)
```

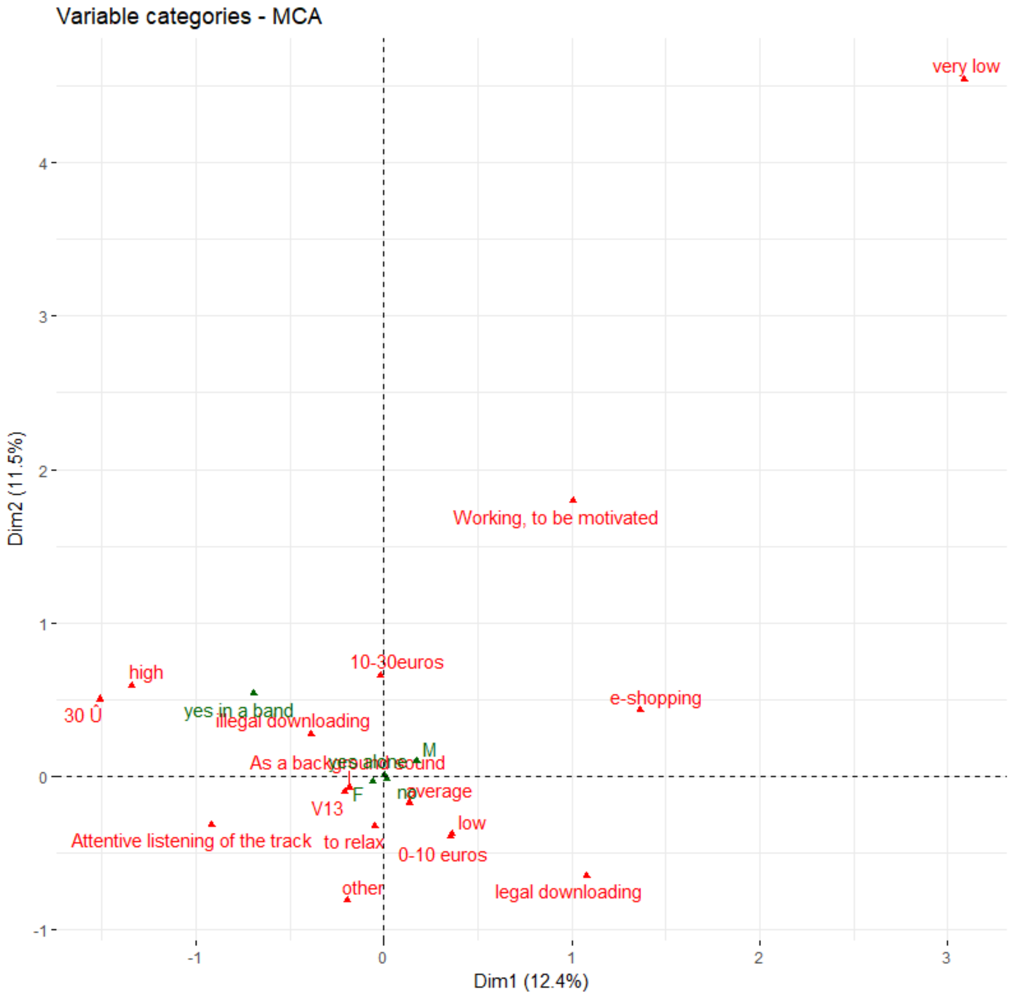


Fig.7

The responses are quite close to the origin; the only big outlier is “very low” which belongs to the knowledge variable

Clustering

The next analysis to carry out is the hierarchical clustering of the individuals with 4 clusters.

```
> res.hcpc<-HCPC(MCA1, nb.clust=4, graph = FALSE)
```

4-clusters hierarchical clustering

```
> MCA2.hcpc<-HCPC(MCA1, nb.clust=4, graph = FALSE)
```

and the plot:

```
plot.HCPC(MCA2.hcpc, axes=c(1,2), choice="map")
```

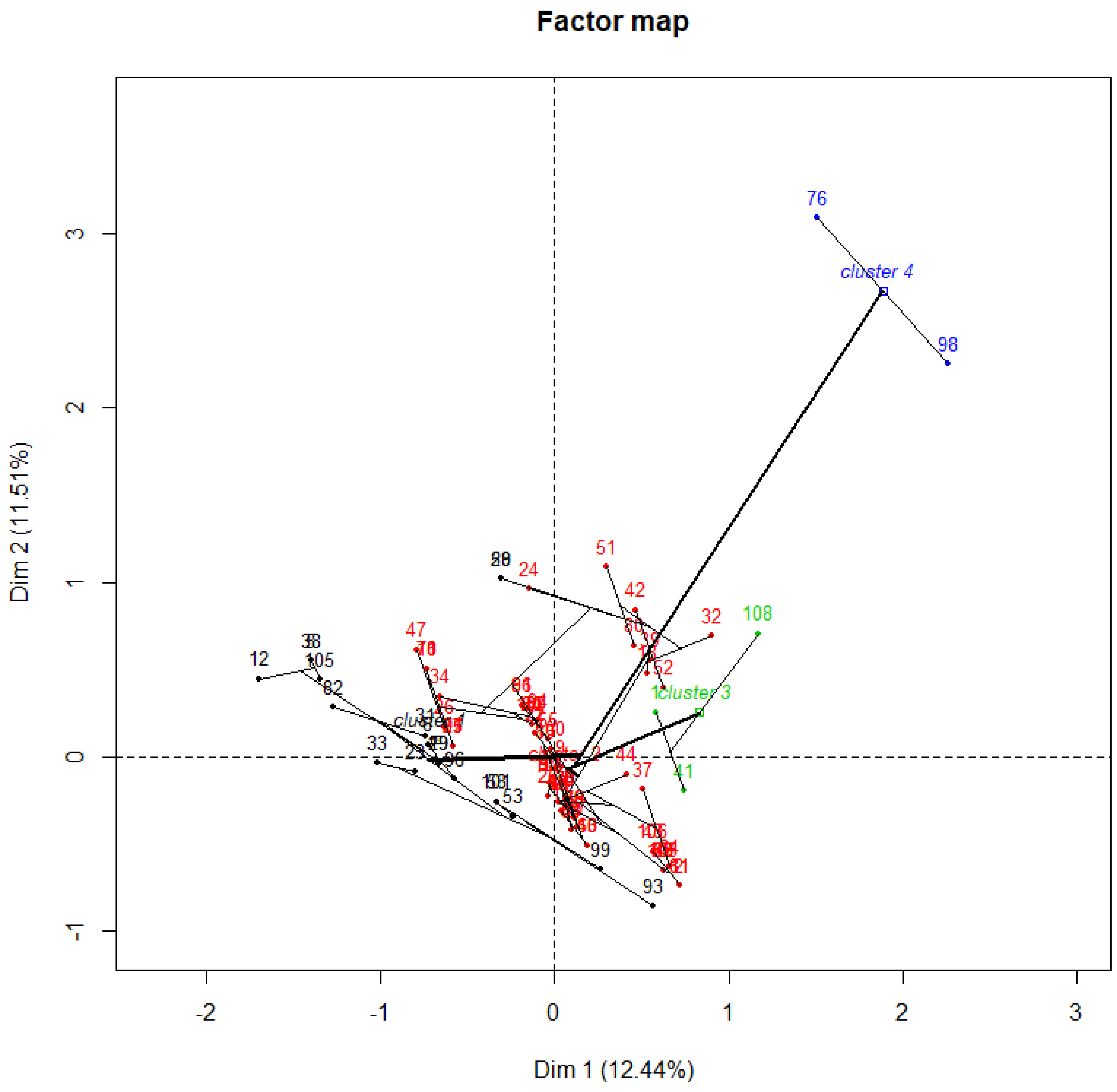


Fig.8

Clusters 4 is detached from the other clusters; while cluster 1 to 3 are next to each other and relatively close to the origin. The clusters are separated but they are quite spread.

Observations

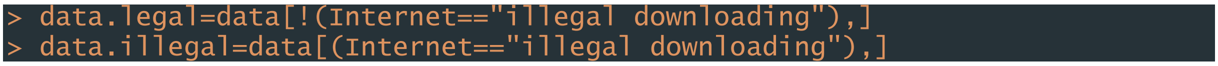
From Fig.3 it is possible to conclude that Knowledge is the main factor that determines behavior; which suggests that the behavior of people with less interest and less musical knowledge is more random and erratic and may be due to other factors not included in the selected variables or the questionnaire as a whole. A larger, more complete and psychologically accurate questionnaire is very necessary in order to extrapolate the information that I was set out to answer. qualitative research is necessary to investigate the reasons behind this trend. This observation is reflected in Fig. 7, where the answer very low in the knowledge question has a big impact.

It can be seen, however (in Fig.5), that the answer “listening as background music” are less impactful than “working, get motivated” which makes logical sense. Another observation is that the answer “30 U”, which means that the respondents wishes to have 30 or more euros of budget for music, as a relatively big impact.

A big limitation of the data is the number of observations, only 108. Moreover, the questions (variables) are pretty simplistic and they do not delve into the psychology. The variable of music genres is quite unhelpful, a better approach wold be a list of preferred genres. Also a question that deals with when people listen to particular genres, for example based on the current mood.

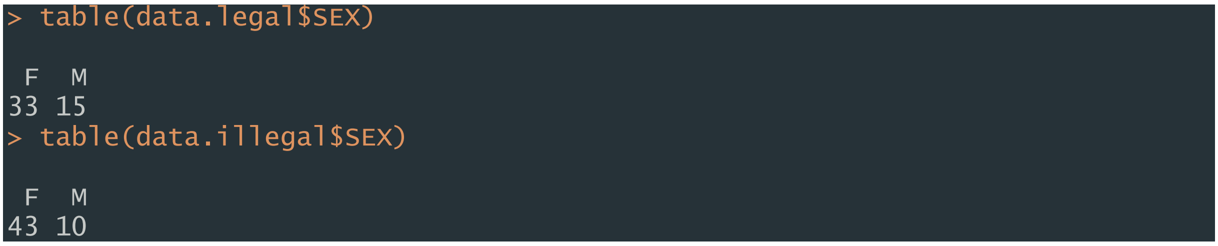
The trend that I set out to find (regarding illegal downloading) is very hard to spot in this analysis and data. It is likely a better approach to find descriptive statistics that show the characteristics of illegal downloaders (e.g. their sex, their budget etc.) which I will do now.

I am going to create two separate datasets, one with the illegal downloads and another without so that a few basic characteristics from those two groups can be identified.



```
> data.legal=data[!(Internet=="illegal downloading"),]
> data.illegal=data[(Internet=="illegal downloading"),]
```

According to the dataset, women are a little bit more inclined to download illegally:



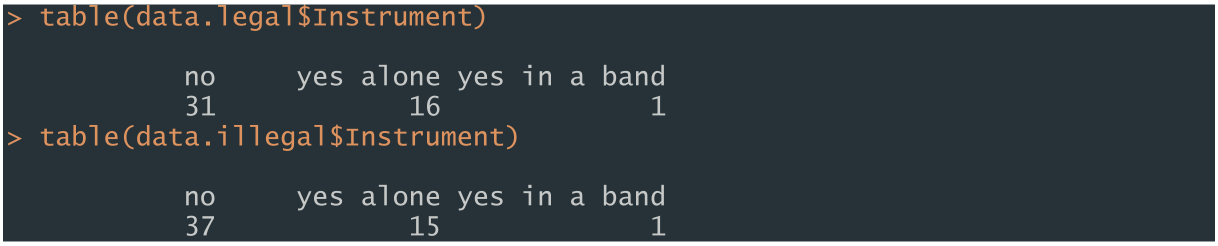
```
> table(data.legal$SEX)

 F  M 
33 15 
> table(data.illegal$SEX)

 F  M 
43 10 
```

There is a total of 76 females and 25 males, females are 304% of the males.
In the legal subset females are only 220% of the males, while in the illegal subset females are 430% the males.

The next command shows how many people play an instrument in the legal and illegal subset:



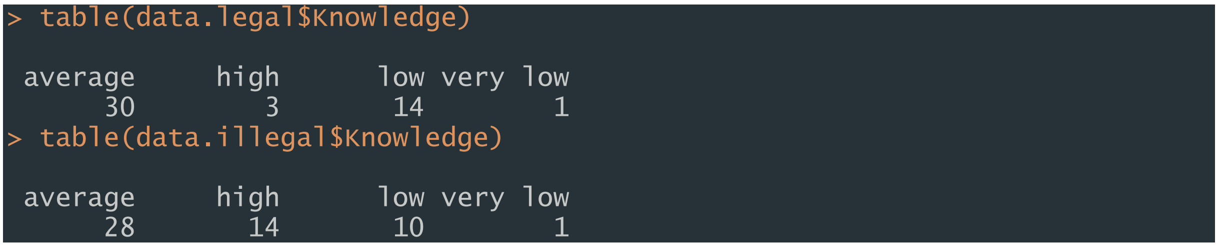
```
> table(data.legal$Instrument)

           no     yes alone yes in a band 
           31            16             1 
> table(data.illegal$Instrument)

           no     yes alone yes in a band 
           37            15             1 
```

It would be logical to think that people who played an instrument are closer to the needs of an artist/musician. But we can see that there is no difference between the two groups.

Now I will explore the knowledge score in the two subsets:



```
> table(data.legal$Knowledge)

 average      high       low very low 
      30         3        14        1 
> table(data.illegal$Knowledge)

 average      high       low very low 
      28        14        10        1 
```

I am going to count “very low” as 0, “low”as 1, “average” as 2 and “high” as 3, add all of the numbers and divide by the number of people in its respective subset.

The “knowledge score” for the legal subset is 1.729 while for the illegal subset is 2.03.
The people that illegally download know more about music than the ones that do not.
A few observations can be made from this: the people that illegally download have access to more music therefore they know more or maybe they are more passionate, require more music but do not have the funds and thus they illegally download. Another interpretation is that the people that do not illegally download are less passionate therefore they did not spend time learning how to safely and easily download music illegally.

A final consideration on this topic is that illegal downloading can be seen in multiple ways, as a pure detriment, a necessary evil or even as advertising for the companies; however, a very good way to deal with this phenomenon is the existence of cheap, easy and streamlined services such as Spotify or other streaming sites/services. Creating platform that people actively prefer to illegal downloading is undeniably better than using invasive DRMs (Digital rights management) tactics and/or law enforcement as it reduces crime and generates revenue at the same time. In fact, streaming services have massively dropped illegal downloading since the previous decade, when such platforms did not exist or were in an embryonal stage (Bernal,2018).

Gastone Riccardo Tolli
ID: 33506151

Date: 29/08/2018

Bibliography

Murtagh F. and Heck A. (1987), Multivariate Data Analysis. D. Reidel Publishing Company

Bernal,N.,2018. Drop in illegal music downloads as streaming sites take over,*Telegraph,*[online]2 August. Available at: https://www.telegraph.co.uk/technology/2018/08/02/drop-illegal-music-downloads-streaming-sites-take/ [Accessed 29 August 2018].