visualisations.R

theo

2022-05-27

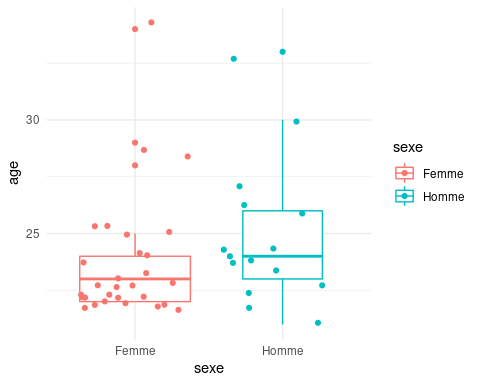
# Packages-------  
  
library(readxl)  
  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.7 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(questionr)  
  
library(FactoMineR)  
  
library(GDAtools)  
  
library(knitr)  
  
# Import & Recodages-----  
  
d <- read\_excel("data/pop\_enq.xlsx") %>%  
   
 mutate(pcs\_pere = fct\_recode(csp\_pere,  
 "Agriculteur" = "1",  
 "Artisan" = "2",  
 "Cadre" = "3",  
 "Profession intermédiaire" = "4",  
 "Employé" = "5",  
 "Ouvrier" = "6",  
 "Autre personne sans activité" = "8",  
 NULL = "NA"),  
 pcs\_mere = fct\_recode(csp\_agr\_mere,  
 "Artisane" = "2",  
 "Cadre" = "3",  
 "Profession intermédiaire" = "4",  
 "Employée" = "5",  
 "Ouvrière" = "6",  
 "Autre personne sans activité" = "8",  
 NULL = "NA"),  
 parent\_cadre = if\_else(csp\_pere == "3" | csp\_agr\_mere == "3",  
 "Au moins un parent cadre",  
 "Aucun parent cadre"))  
  
# Sexe et âge----  
  
ggplot(d, aes(x = sexe, y = age, color = sexe), alpha = 0.7) +  
   
 geom\_boxplot(varwidth = TRUE) +  
   
 geom\_jitter() +  
   
 theme\_minimal()



# PCS des parents--------  
  
table(d$csp\_pere, d$csp\_agr\_mere) %>%  
   
 prop()

##   
## 2 3 4 5 6 8 NA Total  
## 1 0.0 2.3 0.0 0.0 0.0 0.0 0.0 2.3  
## 2 0.0 0.0 7.0 0.0 0.0 0.0 0.0 7.0  
## 3 2.3 18.6 9.3 2.3 2.3 4.7 0.0 39.5  
## 4 0.0 14.0 7.0 4.7 0.0 0.0 0.0 25.6  
## 5 0.0 0.0 0.0 7.0 0.0 0.0 0.0 7.0  
## 6 0.0 0.0 4.7 2.3 0.0 0.0 0.0 7.0  
## 8 0.0 2.3 0.0 0.0 2.3 0.0 0.0 4.7  
## NA 0.0 0.0 0.0 4.7 0.0 0.0 2.3 7.0  
## Total 2.3 37.2 27.9 20.9 4.7 4.7 2.3 100.0

table(d$parent\_cadre) %>%  
   
 freq(digits = 0, valid = FALSE)

## n %  
## Au moins un parent cadre 25 58  
## Aucun parent cadre 18 42

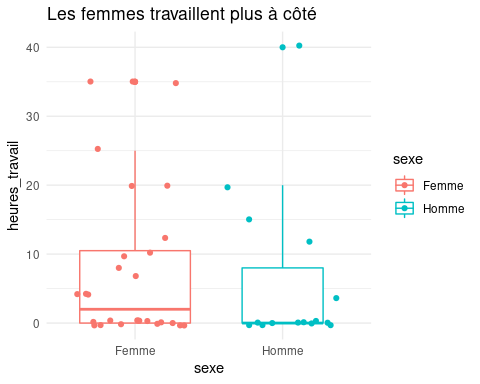
# Origine scolaire et sociale----  
  
table(d$filiere\_bac, d$csp\_pere) %>%  
   
 cprop(digits = 0)

##   
## 1 2 3 4 5 6 8 NA Ensemble  
## Autre 0 33 0 0 33 0 0 0 5   
## ES 100 33 41 33 0 33 50 100 41   
## L 0 0 18 25 33 33 50 0 20   
## Professionnel 0 0 0 17 0 0 0 0 5   
## S 0 33 41 25 33 0 0 0 27   
## STI2D 0 0 0 0 0 33 0 0 2   
## Total 100 100 100 100 100 100 100 100 100

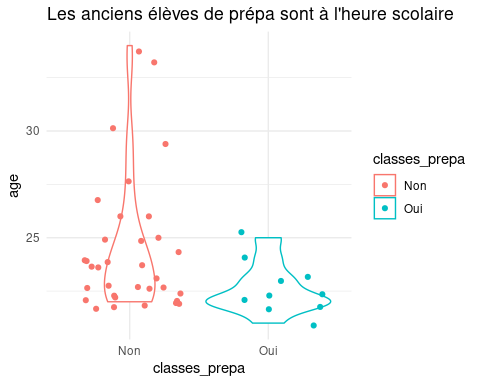
# Sexe et temps de travail-----  
  
ggplot(d, aes(x = sexe, y = heures\_travail, color = sexe), alpha = 0.7) +  
   
 geom\_boxplot(varwidth = TRUE) +  
   
 geom\_jitter() +  
   
 ggtitle("Les femmes travaillent plus à côté") +  
   
 theme\_minimal()

## Warning: Removed 1 rows containing non-finite values (stat\_boxplot).

## Warning: Removed 1 rows containing missing values (geom\_point).

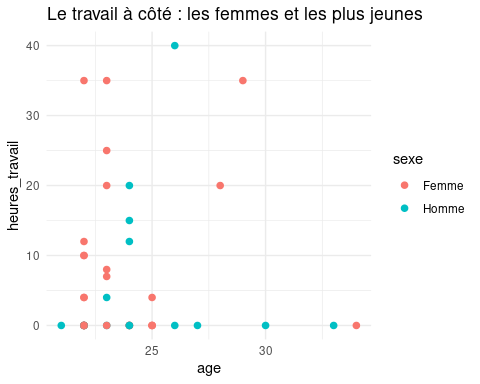


# Classe prépa et âge-------  
  
ggplot(d, aes(x = classes\_prepa, y = age, color = classes\_prepa), alpha = 0.7) +  
   
 geom\_violin() +  
   
 geom\_jitter() +  
   
 ggtitle("Les anciens élèves de prépa sont à l'heure scolaire") +  
   
 theme\_minimal()

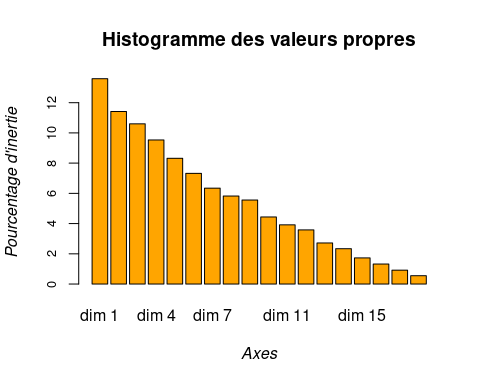


# Age, temps de travail et sexe-------  
  
ggplot(d, aes(x = age, y = heures\_travail, colour = sexe)) +  
   
 geom\_point(size = 2) +  
   
 ggtitle("Le travail à côté : les femmes et les plus jeunes") +  
   
 theme\_minimal()

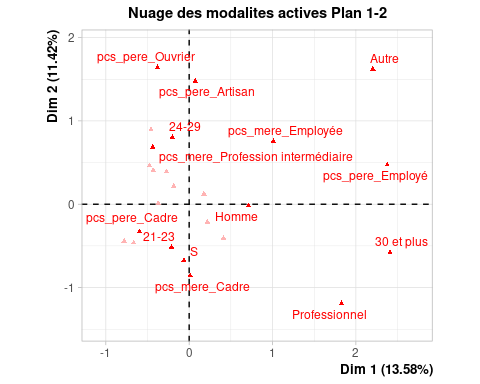
## Warning: Removed 1 rows containing missing values (geom\_point).



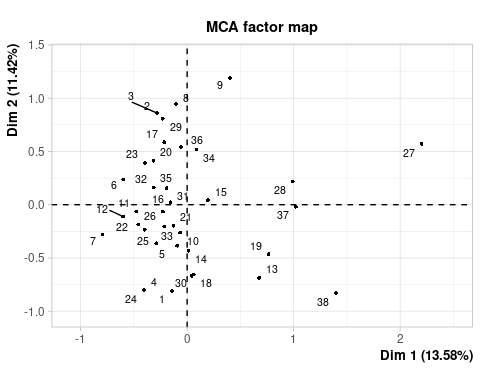
# ACM--------  
  
## Réalisation-----  
  
d\_acm = d %>%  
   
 drop\_na() %>% # on perd 6 individus, ça va encore  
   
 select(tuteur\_hdr, sexe, age\_cat, classes\_prepa, travail, pcs\_pere, pcs\_mere,  
 filiere\_bac) %>%  
   
 modify\_if(is.character, as.factor) # ou : as.data.frame(lapply(d\_acm, factor))  
   
res\_acm = MCA(d\_acm, quali.sup = 1, level.ventil = 0.03, graph = FALSE)  
  
## Valeurs propres---------  
  
# Combien d'axes conserver ?  
  
barplot(res\_acm$eig[,2], main="Histogramme des valeurs propres",   
 xlab="Axes", ylab="Pourcentage d'inertie", cex.axis=0.8, font.lab=3,  
 col="orange")



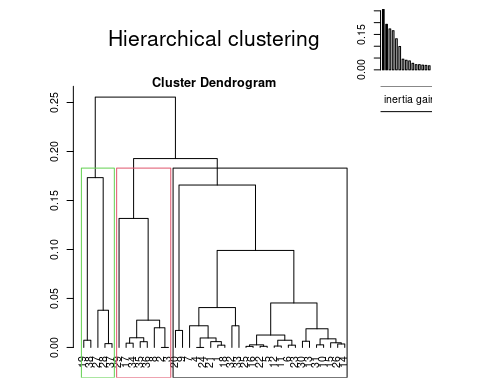
# Le saut est après le 3e on dirait...  
  
## Nuage des modalités les plus contributives-----  
  
seuil <- 100/nrow(res\_acm$var$contrib)  
  
moda <- which(res\_acm$var$contrib[, 1]>seuil   
 | res\_acm$var$contrib[, 2]>seuil)  
  
plot.MCA(res\_acm, invisible=c("ind","quali.sup"),   
 title="Nuage des modalites actives Plan 1-2", axes=c(1,2),  
 selectMod=c(moda), cex=0.8)



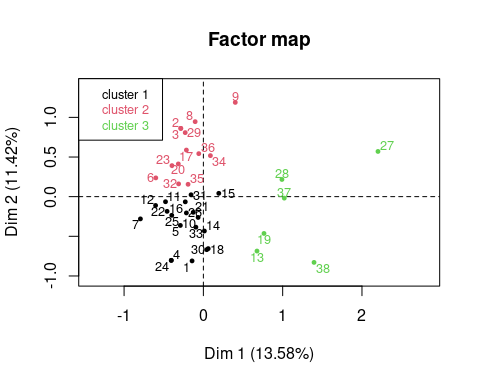
## Nuage des individus---------  
  
plot(res\_acm, choix=c("ind"), invisible=c("var", "quali.sup"), cex=0.7, axes=c(1,2))



# CAH-------  
  
## Réalisation------  
  
res\_cah <- HCPC(res\_acm, nb.clust = 3, graph = FALSE)  
  
## Visualisation des résultats---------  
  
plot(res\_cah, choice = "tree")



plot(res\_cah, choice = "map", draw.tree = F)



## Description des clusters----------  
  
# Quelles sont les variables et modalités qui contribuent le plus aux différents clusters ?  
  
res\_cah$desc.var

##   
## Link between the cluster variable and the categorical variables (chi-square test)  
## =================================================================================  
## p.value df  
## age\_cat 3.699029e-08 4  
## pcs\_pere 3.843300e-04 10  
## pcs\_mere 5.486035e-03 8  
## filiere\_bac 1.604952e-02 8  
## sexe 1.865986e-02 2  
## travail 4.988536e-02 2  
##   
## Description of each cluster by the categories  
## =============================================  
## $`1`  
## Cla/Mod Mod/Cla Global p.value v.test  
## age\_cat=21-23 85.00000 89.473684 52.63158 0.0000051294 4.559424  
## pcs\_mere=pcs\_mere\_Cadre 75.00000 63.157895 42.10526 0.0113314746 2.532307  
## pcs\_pere=pcs\_pere\_Cadre 70.58824 63.157895 44.73684 0.0284956278 2.190395  
## pcs\_mere=pcs\_mere\_Employée 14.28571 5.263158 18.42105 0.0488329238 -1.970048  
## age\_cat=24-29 13.33333 10.526316 39.47368 0.0003289441 -3.591370  
##   
## $`2`  
## Cla/Mod Mod/Cla Global  
## age\_cat=24-29 80.00000 92.307692 39.473684  
## pcs\_mere=pcs\_mere\_Profession intermédiaire 63.63636 53.846154 28.947368  
## pcs\_pere=pcs\_pere\_Ouvrier 100.00000 23.076923 7.894737  
## travail=travaille 58.33333 53.846154 31.578947  
## travail=travaille\_pas 23.07692 46.153846 68.421053  
## pcs\_mere=pcs\_mere\_Cadre 6.25000 7.692308 42.105263  
## age\_cat=21-23 5.00000 7.692308 52.631579  
## p.value v.test  
## age\_cat=24-29 1.971394e-06 4.756335  
## pcs\_mere=pcs\_mere\_Profession intermédiaire 2.332796e-02 2.268020  
## pcs\_pere=pcs\_pere\_Ouvrier 3.390232e-02 2.121232  
## travail=travaille 4.697982e-02 1.986482  
## travail=travaille\_pas 4.697982e-02 -1.986482  
## pcs\_mere=pcs\_mere\_Cadre 2.094419e-03 -3.076506  
## age\_cat=21-23 7.173030e-05 -3.970472  
##   
## $`3`  
## Cla/Mod Mod/Cla Global p.value v.test  
## pcs\_pere=pcs\_pere\_Employé 100.00000 50.00000 7.894737 0.002370792 3.039363  
## age\_cat=30 et plus 100.00000 50.00000 7.894737 0.002370792 3.039363  
## sexe=Homme 38.46154 83.33333 34.210526 0.012897905 2.486576  
## filiere\_bac=Professionnel 100.00000 33.33333 5.263158 0.021337127 2.301965  
## pcs\_pere=pcs\_pere\_Cadre 0.00000 0.00000 44.736842 0.019656020 -2.332850  
## sexe=Femme 4.00000 16.66667 65.789474 0.012897905 -2.486576

# Groupe 1 : enfants de cadre à l'heure scolairement  
# Groupe 2 : enfants d'ouvriers et PI qui travaillent, légèrement "en retard"  
# Groupe 3 : plus âgés, en reprise d'études, hommes, père cadre ou employé  
  
# Quels sont les individus qui sont proches du centre d’un cluster et loin du centre des autres clusters ?  
  
res\_cah$desc.ind

## $para  
## Cluster: 1  
## 5 25 22 16 12   
## 0.1756292 0.3289739 0.3540579 0.4061171 0.4965321   
## ------------------------------------------------------------   
## Cluster: 2  
## 17 23 36 6 35   
## 0.3675590 0.4233811 0.5306166 0.6041039 0.6419615   
## ------------------------------------------------------------   
## Cluster: 3  
## 13 37 28 19 38   
## 0.970236 1.024474 1.195329 1.284423 1.389631   
##   
## $dist  
## Cluster: 1  
## 1 24 4 7 33   
## 1.577607 1.563215 1.563215 1.543898 1.528458   
## ------------------------------------------------------------   
## Cluster: 2  
## 29 9 8 20 3   
## 2.602614 2.404473 1.683133 1.502107 1.471502   
## ------------------------------------------------------------   
## Cluster: 3  
## 27 38 28 19 37   
## 2.612820 2.186842 1.719764 1.687949 1.579242

# Quelles sont les variables qui contribuent le plus à la classification ?  
  
res\_cah$desc.axes

##   
## Link between the cluster variable and the quantitative variables  
## ================================================================  
## Eta2 P-value  
## Dim.1 0.7421848 4.987672e-11  
## Dim.2 0.6244011 3.611401e-08  
##   
## Description of each cluster by quantitative variables  
## =====================================================  
## $`1`  
## v.test Mean in category Overall mean sd in category Overall sd  
## Dim.1 -2.463909 -0.2394066 1.857980e-17 0.2437503 0.5910337  
## Dim.2 -3.816935 -0.3400073 -4.309418e-17 0.2747444 0.5418441  
## p.value  
## Dim.1 0.0137430927  
## Dim.2 0.0001351197  
##   
## $`2`  
## v.test Mean in category Overall mean sd in category Overall sd  
## Dim.2 4.77635 0.5900216 -4.309418e-17 0.3106934 0.5418441  
## p.value  
## Dim.2 1.785059e-06  
##   
## $`3`  
## v.test Mean in category Overall mean sd in category Overall sd  
## Dim.1 5.235761 1.174872 1.85798e-17 0.512926 0.5910337  
## p.value  
## Dim.1 1.643066e-07

# Fréquence par clusters   
  
freq(res\_cah$data.clust$clust)

## n % val%  
## 1 19 50.0 50.0  
## 2 13 34.2 34.2  
## 3 6 15.8 15.8

# Intégrer une variable qualitative "cluster" dans la base de données d'origine   
#qui precise dans quel cluster se situe l'individu   
  
d\_acm$cluster <- res\_cah$data.clust$clust