Pràctica part 2: projecte de visualització

Arnau Janot Baró (ajanot)

2024-01-13

Conjunt de dades seleccionat

Per a la realització de la pràctica final treballarem amb el dataset netejat en la part 1 de la pràctica *gpa_clean.csv* que conté la nota mitjana d'estudiants universitaris després del primer semestre de classes (*GPA*: grade point average, en anglès), així com informació sobre la nota d'accés, la cohort de graduació a l'institut i algunes característiques dels estudiants.

Aquest conjunt de dades surt d'una enquesta realitzada a una mostra representativa d'estudiants d'una universitat dels EUA.

Les variables incloses al conjunt de dades són les següents:

• **sat: (num)** nota d'accés (escala de 400 a 1600 punts)

• **tothrs: (num)** hores totals cursades al semestre

• **hsize: (int)** nombre total d'estudiants a la cohort de graduats del batxillerat (en centenars)

• **hsrank: (num)** rànquing de l'estudiant, donat per la nota mitjana del batxillerat, en la cohort de graduats del batxillerat

• **hsperc: (num)** rànquing relatiu de l'estudiant (hsrank/hsize)

• **colgpa: (factor)** nota mitjana de l'estudiant al final del primer semestre (escala de 0 a 4 punts)

• **athlete: (factor)** indicador de si l'estudiant practica algun esport a la universitat

• **female: (factor)** indicador de si l'estudiant és dona

• **white: (factor)** indicador de si l'estudiant és de raça blanca o no

• **black: (factor)** ndicador de si l'estudiant és de raça negra o no

• **gpaletter: (factor)** indicador dels resultats en forma de lletra segons la nota

Realitzarem un anàlisi entorn a estudiar la nota dels estudiants a partir de les variables d'interès així com la proporció d'atletes entre la població d'estudiants. És interessant veure si aspectes com el fet de practicar esport pot influir en les notes a

l'escola, o inclus si aspectes com el sexe tenen una influència a l'escola en termes de resultats.

Ens plantejem les següents preguntes:

• Ser atleta influeix a la nota? Si un estudiant practica esport a la universitat obtindra millors resultats?

• Les dones obtenen millor nota que els homes?

• Les persones de raça negra obtenen pitjors resultats que les persones que no son negres?

• Hi ha més atletes entre els homes que entre les dones?

Per tal de respondre a aquestes preguntes crearem visualitzacions respecte a les dades que disposem. Les visualitzacions és crearan a través de Rstudio i de tableau.

Primer de tot carreguem les dades a Rstudio i les mostrem en pantalla.

```
gpa <- read.csv("gpa_clean.csv")
head(gpa)
```

```
##    sat tothrs hsize hsrank hsperc colgpa athlete female white black
gpaletter
## 1  920     43  0.10      4 40.000   2.04    TRUE   TRUE FALSE FALSE
C
## 2 1170     18  4.86    191 20.319   4.00   FALSE  FALSE  TRUE FALSE
A
## 3  810     14  1.19     42 35.294   1.78    TRUE  FALSE  TRUE FALSE
C
## 4  940     40  5.71    252 44.133   2.42   FALSE  FALSE  TRUE FALSE
C
## 5 1180     18  2.14     86 40.187   2.61   FALSE  FALSE  TRUE FALSE
B
## 6  980    114  2.68     41 15.299   3.03   FALSE   TRUE  TRUE FALSE
B
```

Ara examinarem el tipus de dades que tenim en el nostre dataset.

```
str(gpa)
```

```
## 'data.frame':    4137 obs. of  11 variables:
##  $ sat      : int  920 1170 810 940 1180 980 880 980 1240 1230 ...
##  $ tothrs   : int  43 18 14 40 18 114 78 55 18 17 ...
##  $ hsize    : num  0.1 4.86 1.19 5.71 2.14 ...
##  $ hsrank   : int  4 191 42 252 86 41 161 101 161 3 ...
##  $ hsperc   : num  40 20.3 35.3 44.1 40.2 ...
##  $ colgpa   : num  2.04 4 1.78 2.42 2.61 ...
##  $ athlete  : logi  TRUE FALSE TRUE FALSE FALSE FALSE ...
##  $ female   : logi  TRUE FALSE FALSE FALSE FALSE TRUE ...
##  $ white    : logi  FALSE TRUE TRUE TRUE TRUE TRUE ...
```

```
##  $ black    : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ gpaletter: chr  "C" "A" "C" "C" ...
```

Com podem veure, el dataset conté actualment 4137 registres, 11 variables de tipus int, num, logical i chr.

Anem ara a fer una primera aproximació sobre les dades que tenim amb la funció summary. La funció summary de R ens proporciona un resum estadístic per a cada variable. Mostra el mínim, el primer quartil, la mediana, la mitjana, el tercer quartil i el màxim.

```
summary(gpa)
```

```
##       sat           tothrs          hsize           hsrank      
##  Min.   : 700   Min.   :  6.00   Min.   :0.030   Min.   :  1.00  
##  1st Qu.: 940   1st Qu.: 17.00   1st Qu.:1.650   1st Qu.: 11.00  
##  Median :1030   Median : 47.00   Median :2.500   Median : 30.00  
##  Mean   :1031   Mean   : 52.83   Mean   :2.611   Mean   : 52.83  
##  3rd Qu.:1120   3rd Qu.: 80.00   3rd Qu.:3.500   3rd Qu.: 70.00  
##  Max.   :1390   Max.   :137.00   Max.   :6.000   Max.   :634.00  
##      hsperc           colgpa       athlete          female      
##  Min.   : 0.167   Min.   :0.000   Mode :logical   Mode :logical  
##  1st Qu.: 6.433   1st Qu.:2.210   FALSE:3943      FALSE:2277     
##  Median :14.583   Median :2.660   TRUE :194       TRUE :1860     
##  Mean   :19.237   Mean   :2.654                                  
##  3rd Qu.:27.711   3rd Qu.:3.120                                  
##  Max.   :92.000   Max.   :4.000                                  
##    white            black          gpaletter        
##  Mode :logical   Mode :logical   Length:4137        
##  FALSE:308       FALSE:3908      Class :character   
##  TRUE :3829      TRUE :229       Mode  :character   
##                                                     
##                                                     
##
```

Visualització

Estudiarem de forma visual la distribució de la variable *colgpa* (nota mitjana de l'estudiant al final del primer semestre (escala de 0 a 4 punts)) en funció de les variables d'interés female, athlete i black per tal de poder respondre a les preguntes que ens hem p`lantejat al principi.

Anem a mostrar la distribució de la variable 'colgpa' respecte a la variable gènere ('female'), la variable atleta ('athlete') i la raça ('black')

Visualització de les variables sat i colgpa en funció de les variables female, athlete i black.

```
# Crearem els objectes que representin la mitjana de cada grup (TRUE i
FALSE)
```

```r
# de cada variable qualitativa (female, athlete i black) per la variable colgpa

mitj_T_F_female_colgpa <- plyr::ddply(gpa, "female", plyr::summarise, grp.mean=mean(colgpa))
head(mitj_T_F_female_colgpa)
```

```
##   female grp.mean
## 1  FALSE 2.589293
## 2   TRUE 2.733059
```

```r
mitj_T_F_athlete_colgpa <- plyr::ddply(gpa, "athlete", plyr::summarise, grp.mean=mean(colgpa))
head(mitj_T_F_athlete_colgpa)
```

```
##   athlete grp.mean
## 1   FALSE 2.667337
## 2    TRUE 2.381443
```

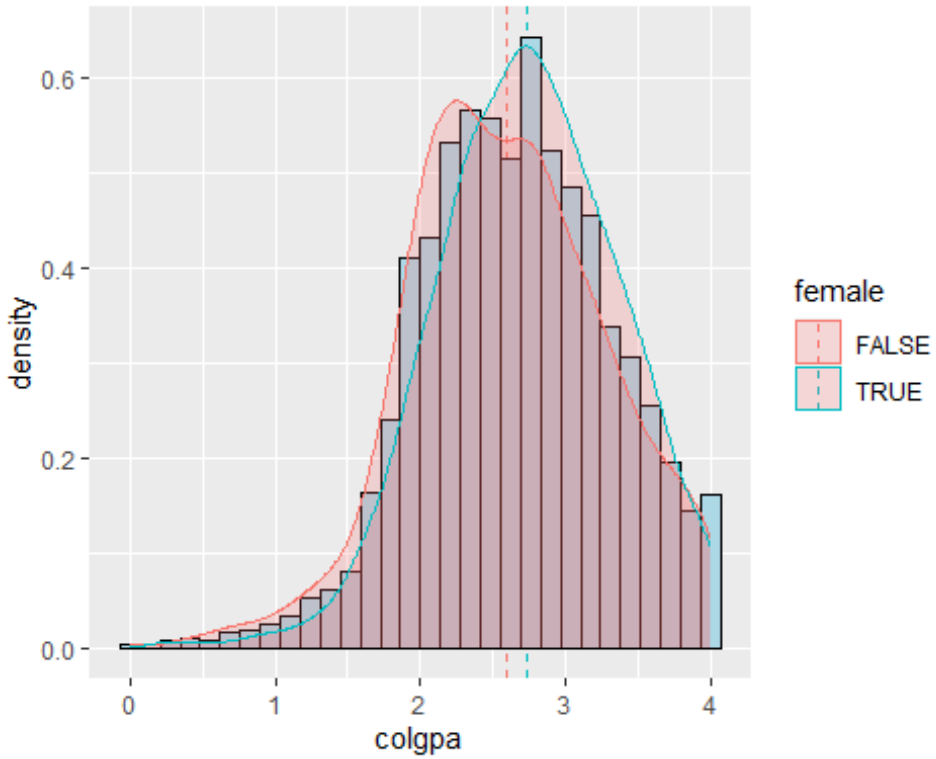
```r
mitj_T_F_black_colgpa <- plyr::ddply(gpa, "black", plyr::summarise, grp.mean=mean(colgpa))
head(mitj_T_F_black_colgpa)
```

```
##   black grp.mean
## 1 FALSE 2.677262
## 2  TRUE 2.255764
```

```r
# Creem ara els gràfics
# colgpa en funció de female
g1 <- ggplot2::ggplot(gpa, ggplot2::aes(x = colgpa, color=female)) +
ggplot2::geom_vline(data=mitj_T_F_female_colgpa,
ggplot2::aes(xintercept=grp.mean, color=female), linetype="dashed") +
ggplot2::geom_histogram(ggplot2::aes(y=..density..),
colour="black", fill="lightblue") +
ggplot2::geom_density(alpha=.2, fill="#FF6666")
g1
```

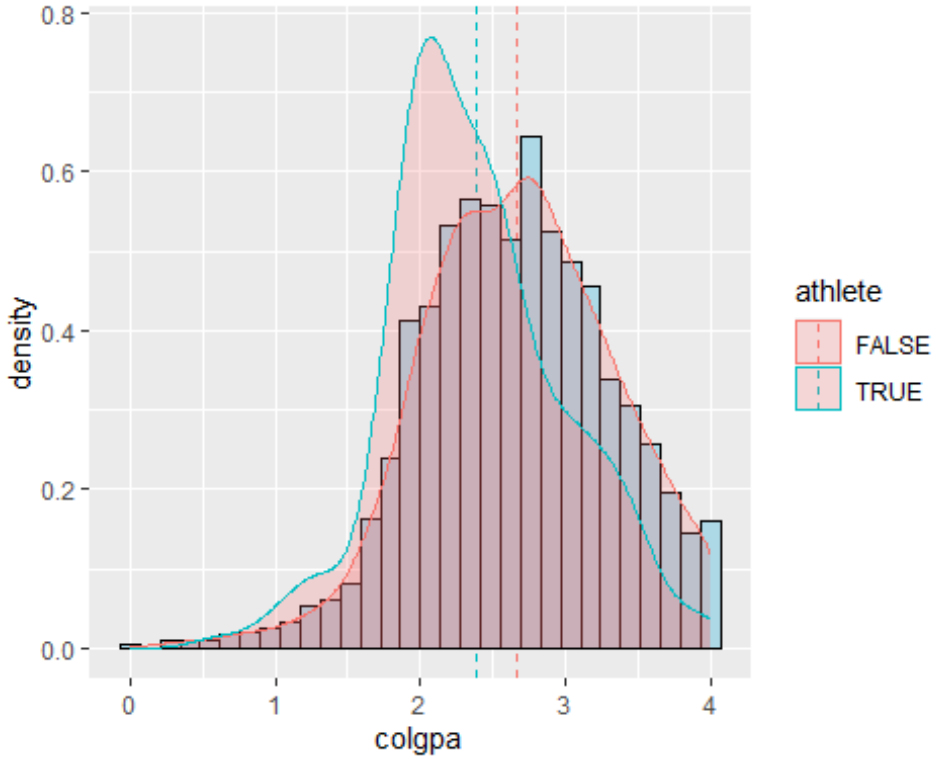
```
## Warning: The dot-dot notation (`..density..`) was deprecated in
ggplot2 3.4.0.
## ℹ Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning
was
## generated.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



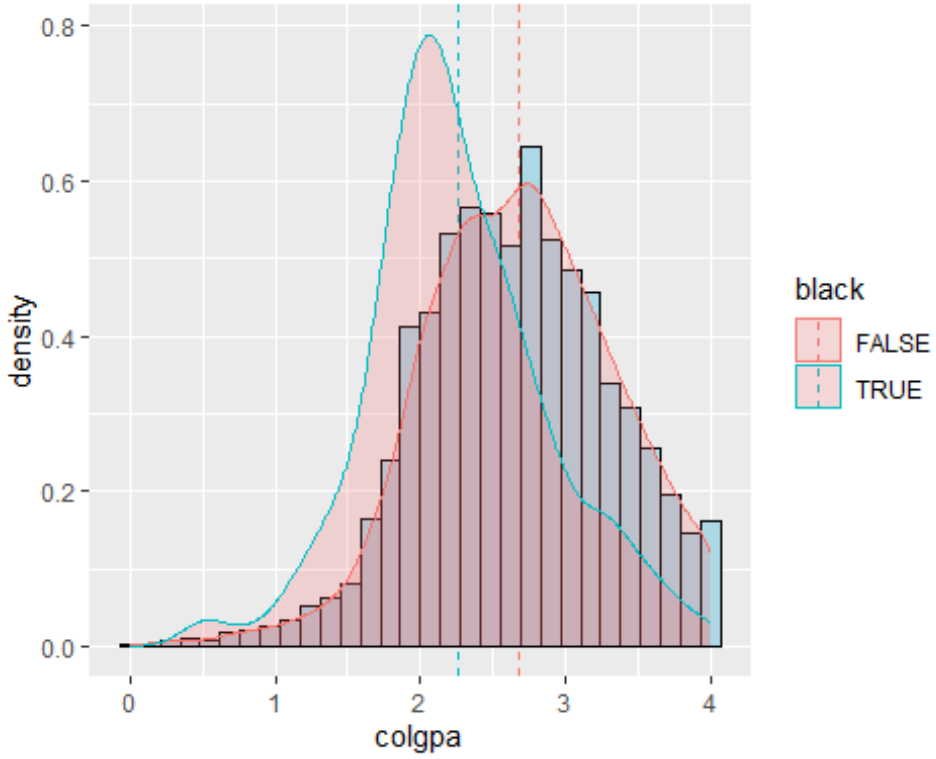

```
# colgpa en funció de athlete
g2 <- ggplot2::ggplot(gpa, ggplot2::aes(x = colgpa, color=athlete)) +
ggplot2::geom_vline(data=mitj_T_F_athlete_colgpa,
ggplot2::aes(xintercept=grp.mean, color=athlete), linetype="dashed") +
ggplot2::geom_histogram(ggplot2::aes(y=..density..),
colour="black", fill="lightblue") +
ggplot2::geom_density(alpha=.2, fill="#FF6666")
g2

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# colgpa en funció de black
g3 <- ggplot2::ggplot(gpa, ggplot2::aes(x = colgpa, color=black)) +
ggplot2::geom_vline(data=mitj_T_F_black_colgpa,
ggplot2::aes(xintercept=grp.mean, color=black), linetype="dashed") +
ggplot2::geom_histogram(ggplot2::aes(y=..density..),
colour="black", fill="lightblue") +
ggplot2::geom_density(alpha=.2, fill="#FF6666")
g3
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Hi ha més atletes entre els homes que entre les dones?

Per tal de poder respondre a aquesta pregunta, crearem ara un gràfic amb la variable athlete en funció del sexe on es mostri visualment si hi ha més homes que fan esport en la universitat que dones.

Convinem primer de tot les variables female i athlete amb la funció table(), i creem la variable que volem.

```
tabla1 <- table(gpa$female, gpa$athlete)
tabla1

##
##         FALSE TRUE
##   FALSE  2128  149
##   TRUE   1815   45

tabla2 <- prop.table(tabla1, margin = 2)
tabla2

##
##              FALSE      TRUE
##   FALSE 0.5396906 0.7680412
##   TRUE  0.4603094 0.2319588
```

Creem ara el gràfic amb les variables tabla1 i tabla2 que hem creat. Utilitzarem la funció barplot().

```
barplot(tabla2, las = 1, xlab = "Atleta",   ylab = "Percentatge del
sexe",
names.arg = c("Atleta no", "Atleta si"), col = c("lightblue",
"mistyrose"),
ylim = c(0, 1), main = "Percentatge de homes i dones atletes o no")
legend("topright", legend = c("Homes", "Dones"), bty = "n",
       fill = c("lightblue", "mistyrose"), title = "Sexe")
```