



# Estudi de l'èxit i l'abandonament acadèmic dels universitaris

Arroyo Luque, Oscar

Bohils Tenas, Mireia

Llaneras Casas, Aina

Martinez Margareto, Lucía

Moyano Núñez, Berta

Ortiz Vallès, Albert

Torrents Alcañiz, Laura

Ventosa Andreu, Laura

Barcelona, 24 de maig de 2023

# Contents

<b>1 Definició de projecte i assignació</b>	<b>3</b>
1.1 Font d'obtenció de les dades . . . . .	3
1.2 Informació de les dades . . . . .	3
1.3 Estructura i informació de la matriu de dades . . . . .	3
1.4 Missings . . . . .	6
<b>2 Pla de treball</b>	<b>8</b>
2.1 Diagrama de Gantt . . . . .	8
2.2 Assignació de tasques . . . . .	9
2.3 Pla de riscos . . . . .	9
<b>3 Estructura de les dades i descriptiva</b>	<b>10</b>
3.1 Selecció de columnes . . . . .	10
3.2 Anàlisi descriptiva abans del preprocessing . . . . .	10
3.2.1 Anàlisi univariant de les variables . . . . .	10
3.2.2 Anàlisi bivariant . . . . .	21
3.3 Preprocessing . . . . .	38
3.3.1 Detecció d'outliers . . . . .	38
3.3.2 Tractament de NA's . . . . .	39
3.4 Anàlisi descriptiva després del preprocessing . . . . .	39
3.4.1 Anàlisi univariant de les variables . . . . .	39
3.4.2 Anàlisi bivariant . . . . .	50
<b>4 Clustering jeràrquic</b>	<b>74</b>
4.1 Clustering . . . . .	74
4.2 Profiling dels clusters . . . . .	76
4.2.1 Anàlisi gràfica de les variables numèriques . . . . .	78
4.2.2 Anàlisi gràfica de les variables categòriques . . . . .	81
4.2.3 Conclusió sobre els clústers a les variables numèriques i categòriques . . . . .	89
<b>5 ACP</b>	<b>90</b>
5.1 Projecció de les variables numèriques . . . . .	93
5.2 Projecció de les variables categòriques . . . . .	96
<b>6 ACM</b>	<b>105</b>
6.1 Plot individus . . . . .	109
6.2 Plot variables . . . . .	110
6.3 Associacions . . . . .	114

<b>7 Clustering jeràrquic sobre les components factorials retingudes a l'ACP</b>	<b>115</b>
7.1 Profiling post ACP i ACM . . . . .	116
7.1.1 Anàlisi gràfica de les variables numèriques . . . . .	117
7.1.2 Anàlisi gràfica de les variables categòriques . . . . .	120
7.1.3 Conclusió sobre els clústers a les variables numèriques i categòriques . . . . .	128
7.1.4 Comparació dels profilings . . . . .	128
7.2 Clústers representats sobre els dos primers plans factorials . . . . .	129
<b>8 Anàlisi discriminant lineal</b>	<b>131</b>
<b>9 Anàlisi textual</b>	<b>135</b>
9.1 Preprocessament . . . . .	135
9.2 Topic modeling . . . . .	136
9.2.1 Per dos tòpics . . . . .	136
9.2.2 Per quatre tòpics . . . . .	137
<b>10 Anàlisi comparativa i conclusió</b>	<b>139</b>
<b>11 Pla de treball real</b>	<b>140</b>
11.1 Diagrama de Gantt . . . . .	140
11.2 Assignació de tasques . . . . .	142
11.3 Pla de riscos . . . . .	143

# 1 Definició de projecte i assignació

## 1.1 Font d'obtenció de les dades

Les dades han estat extretes del repositori de bases de dades Kaggle. L'enllaç de la pàgina web és el següent:  
<https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-studentretention>

Aquí se citen els autors de la base de dades:

*VALORIZA—Research Center for Endogenous Resource Valorization, Instituto Politécnico de Portalegre, 7300-555 Portalegre, Portugal.*

*Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Portalegre, 7300-555 Portalegre, Portugal.*

## 1.2 Informació de les dades

La base de dades inclou diversos casos de dades demogràfiques, socioeconòmiques i macroeconòmiques d'alumnes universitaris, així com els seus resultats acadèmics entre el 2008 i el 2018. Aquests resultats s'enforcaran com a variable resposta, per procurar establir una relació entre l'èxit acadèmic i les altres variables.

Aquestes inclouen dades de 17 graus universitaris de diferents camps de coneixement, com l'agronomia, el disseny, l'educació, la infermeria, el periodisme, la gestió, el servei social i tecnologies.

## 1.3 Estructura i informació de la matriu de dades

Files (individus)	Columnes (variables)	Nre. variables numèriques	Nre. variables categòriques
4424	34	8	10

### INFORMACIÓ DEMOGRÀFICA

Nom	Descripció	Tipus	Diccionari i domini
Nationality	Nacionalitat	Categòrica	1—Portuguese, 2—German, 3—Spanish 4—Italian, 5—Dutch, 6—English, 7—Lithuanian, 8—Angolan, 9—Cape Verdean, 10—Guinean, 11—Mozambican, 12—Santomean, 13—Turkish, 14—Brazilian, 15—Romanian, 16—Moldova, 17—Mexican, 18—Ukrainian, 19—Russian, 20—Cuban, 21—Colombian
Gender	Gènere	Categòrica	1—Male, 0—Female
Age	Edat	Numèrica	Rang a les dades [17, 70]

### DADES SOCIOECONÒMIQUES

Nom	Descripció	Tipus	Diccionari i domini
Scholarship holder	Becari	Categòrica binària	1—Yes, 0—No

<b>Nom</b>	<b>Descripció</b>	<b>Tipus</b>	<b>Diccionari i domini</b>
Mother's occupation & Father's occupation	Ocupació de la mare i Ocupació del pare	Categòrica	1—Student, 2—Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers, 3—Specialists in Intellectual and Scientific Activities, 4—Intermediate Level Technicians and Professions, 5—Administrative staff, 6—Personal Services, Security and Safety Workers, and Sellers, 7—Farmers and Skilled Workers in Agriculture, Fisheries, and Forestry, 8—Skilled Workers in Industry, Construction, and Craftsmen, 9—Installation and Machine Operators and Assembly Workers, 10—Unskilled Workers, 11—Armed Forces Professions, 12—Other Situation, 13—(blank), 14—Armed Forces Officers, 15—Armed Forces Sergeants, 16—Other Armed Forces personnel, 17—Directors of administrative and commercial services, 18—Hotel, catering, trade, and other services directors, 19—Specialists in the physical sciences, mathematics, engineering, and related techniques, 20—Health professionals, 21—Teachers, 22—Specialists in finance, accounting, administrative organization, and public and commercial relations, 23—Intermediate level science and engineering technicians and professions, 24—Technicians and professionals of intermediate level of health, 25—Intermediate level technicians from legal, social, sports, cultural, and similar services, 26—Information and communication technology technicians, 27—Office workers, secretaries in general, and data processing operators, 28—Data, accounting, statistical, financial services, and registry-related operators, 29—Other administrative support staff, 30—Personal service workers, 31—Sellers, 32—Personal care workers and the like, 33—Protection and security services personnel, 34—Market-oriented farmers and skilled agricultural and animal production workers, 35—Farmers, livestock keepers, fishermen, hunters and gatherers, and subsistence, 36—Skilled construction workers and the like, except electricians, 37—Skilled workers in metallurgy, metalworking, and similar, 38—Skilled workers in electricity and electronics, 39—Workers in food processing, woodworking, and clothing and other industries and crafts, 40—Fixed plant and machine operators, 41—Assembly workers, 42—Vehicle drivers and mobile equipment operators, 43—Unskilled workers in agriculture, animal production, and fisheries and forestry, 44—Unskilled workers in extractive industry, construction, manufacturing, and transport, 45—Meal preparation assistants, 46—Street vendors (except food) and street service providers

Nom	Descripció	Tipus	Diccionari i domini
Educational special needs	Necessitats especials en l'educació	Categòrica binària	1—Yes, 0—No

## DADES MACROECONÒMIQUES

Nom	Descripció	Tipus	Diccionari i domini
Unemployment rate	Taxa d'atur	Numèrica	Rang a les dades [7.6, 16.2]
Inflation rate	Taxa d'inflació	Numèrica	Rang a les dades [-0.8 , 3,7]
GDP	PIB	Numèrica	Rang a les dades [-4.1, 3.5]

## DADES ACADÈMIQUES A LA INSCRIPCIÓ

Nom	Descripció	Tipus	Diccionari i domini
Course	Curs	Categòrica	1—Biofuel Production Technologies, 2—Animation and Multimedia Design, 3—Social Service (evening attendance), 4—Agronomy, 5—Communication Design, 6—Veterinary Nursing, 7—Informatics Engineering, 8—Equiniculture, 9—Management, 10—Social Service, 11—Tourism, 12—Nursing, 13—Oral Hygiene, 14—Advertising and Marketing Management, 15—Journalism and Communication, 16—Basic Education, 17—Management (evening attendance)
Attendance	Assistència	Categòrica binària	1—Daytime, 0—Evening
Previous qualification	Qualificació prèvia	Categòrica	1—Secondary education, 2—Higher education—bachelor's degree, 3—Higher education—degree, 4—Higher education—master's degree, 5—Higher education—doctorate, 6—Frequency of higher education, 7—12th year of schooling—not completed, 8—11th year of schooling—not completed, 9—Other—11th year of schooling, 10—10th year of schooling, 11—10th year of schooling—not completed, 12—Basic education 3rd cycle (9th/10th/11th year) or equivalent, 13—Basic education 2nd cycle (6th/7th/8th year) or equivalent, 14—Technological specialization course, 15—Higher education—degree (1st cycle), 16—Professional higher technical course, 17—Higher education—masters degree (2nd cycle)

## DADES ACADÈMIQUES AL FINAL DEL 1r SEMESTRE

Nom	Descripció	Tipus	Diccionari i domini
Curricular units 1st sem grades	Unitats curriculars 1r semestre notes	Numèrica	Rang a les dades [0, 18.875]
Curricular units 1st sem without evaluations	Unitats curriculars 1r semestre sense evaluacions	Numèrica	Rang a les dades [0, 12]

## DADES ACADÈMIQUES AL FINAL DEL 2n SEMESTRE

Nom	Descripció	Tipus	Diccionari i domini
Curricular units 2nd sem grades	Unitats curriculars 2n semestre notes	Numèrica	Rang a les dades [0, 18.571]
Curricular units 2nd sem without evaluations	Unitats curriculars 2n semestre sense evaluacions	Numèrica	Rang a les dades [0, 12]

## VARIABLE D'OUTPUT

Nom	Descripció	Tipus	Diccionari i domini
Target	Target	Categòrica	Graduate, Dropout, Enrolled

## 1.4 Missings

Nre. de caselles	Respecte del total de la matriu dades
120	0.0775%

```
## Mother.s.qualification Father.s.qualification
##                      0.1808                  2.5316
```



Al gràfic només s'han representat les variables que tenen algun missing.

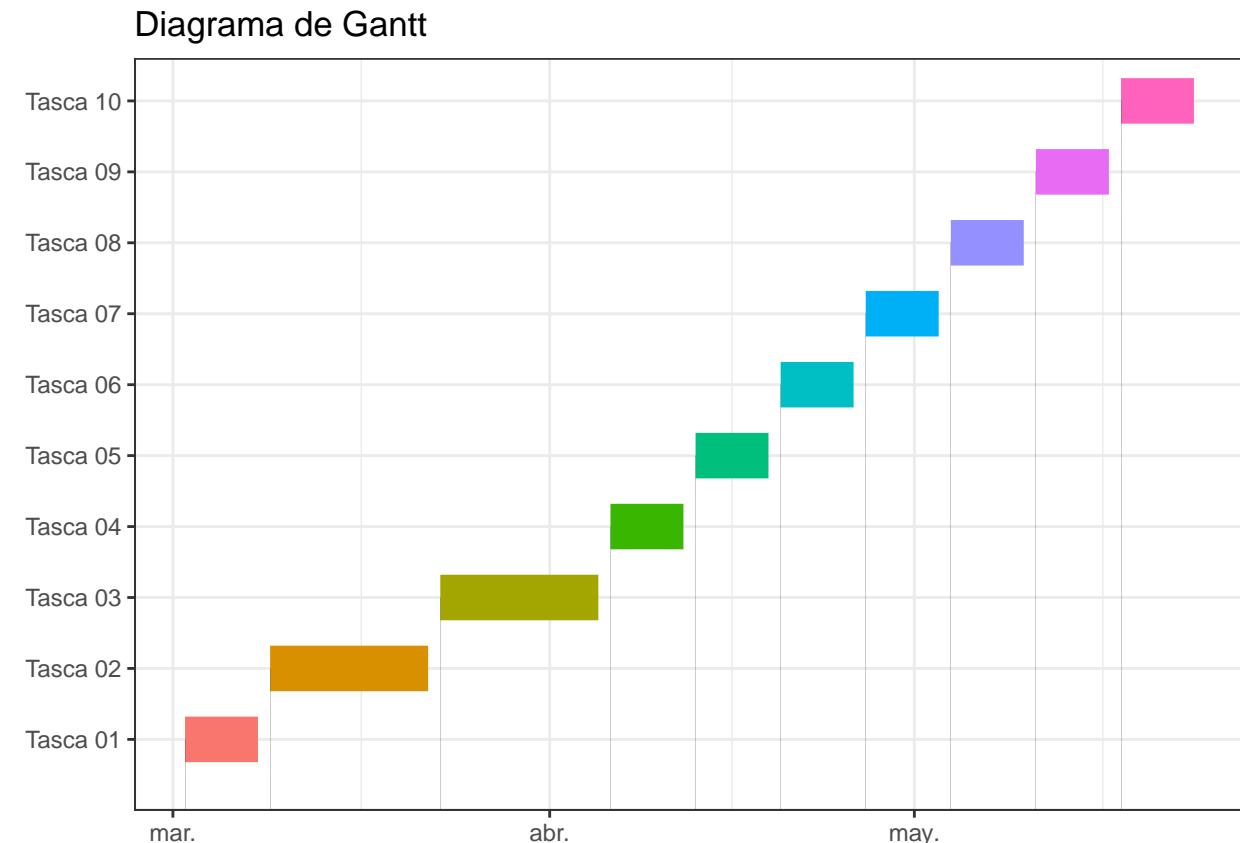
## 2 Pla de treball

### 2.1 Diagrama de Gantt

Per tal de complir els objectius del treball i seguir tots els punts que consten en el guió, s'ha dividit la feina en les següents tasques a desenvolupar.

1. Motivació del treball, descripció formal de les dades i profiling de les dades crues
2. Preprocessament
3. Profiling dades preprocessades i clustering jeràrquic
4. Preparar informe D3 i presentació 1
5. ACP i ACM
6. Clustering jeràrquic de l'ACP i l'ACM
7. Anàlisi comparativa
8. Conclusions
9. Pla de treball real
10. Preparar informe D4 i presentació 2

Aquesta és la planificació de quan es duran a terme cadascuna de les tasques.



## 2.2 Assignació de tasques

Les tasques descrites en l'anterior secció s'han repartit entre els integrants del grup de la següent manera:

Tasca	Aina i Oscar	Lucía i Laura T.	Mireia i Laura V.	Berta i Albert
Motivació treball	X			
Descripció formal dades		X		
Profiling dades crues			X	
Preprocessament				X
Profiling dades preprocessades		X		
Clustering jeràrquic	X			
Informe D3 i presentació 1	X	X	X	X
ACP			X	
ACM				X
Clustering jeràrquic	X			
ACP i ACM				
Anàlisi comparativa		X		
Conclusions	X	X	X	X
Pla treball real				X
Informe D4 i presentació 2	X	X	X	X

## 2.3 Pla de riscos

Com prèviament s'ha indicat, s'ha organitzat els temps de durada i l'ordre de cada tasca en un diagrama de Gantt. Així, amb la premissa de prevenir que qualsevol imprevist s'agreugi i poder continuar aquesta planificació, s'han pactat els següents punts com a pla de riscos:

Risc	Impacte	Solució
Falta de temps acadèmic, retard en les tasques	Alt	Planificació amb suficient antelació. Si cal, reajustar el calendari, allargar i escurçar terminis.
Pèrdua de documents	Alt	Possible recuperació de versions anteriors. Tenir còpies de totes les actualitzacions (Drive i GitHub).
Renúncia o pèrdua de l'avaluació contínua	Moderat	Es reorganitzen els grups.
Incapacitat per treballar (indisposició, malalties, problemes personals, problemes informàtics, etc.)	Moderat	Si la tasca té poc pes, la fa la parella. Si és molta feina, s'encarrega un altre grup que no tingui cap tasca en el moment.
Conflictes entre parelles	Baix	Intentar solucionar els problemes amb ajuda d'algun membre com a mediador. Si no s'aconsegueix, reestructuració dels grups.

L'impacte fa referència al nivell del risc i, per tant, la prioritat que se li ha de donar a aquest.

### 3 Estructura de les dades i descriptiva

#### 3.1 Selecció de columnes

Abans de començar a explorar les dades i treballar sobre elles se seleccionen les columnes que més ens interessen de la base de dades escollida pel treball. Les columnes se seleccionen de tal manera que compleixin els requisits plantejats, un mínim de 7 columnes numèriques i un mínim de 7 categòriques, on almenys 2 són binàries. Les columnes amb les quals ens quedem s'han explicat en la secció anterior d'aquest informe.

Per tal de facilitar la interpretació de les dades al llarg del treball, també es recodifiquen les variables categòriques amb el valor que correspon a cada nivell. Per exemple, per a la variable binària *daytime\_attendance*, s'assigna el valor *Daytime* als 0's i *Evening* als 1's, d'acord amb el diccionari que accompanyava la base de dades. D'aquesta manera, els diferents nivells de les variables categòriques poden interpretar-se directament.

També s'arrodoneixen a dos decimals els valors de les variables numèriques per facilitar la comprensió i simplificar les dades a l'hora de treballar. Arrodonir valors numèrics no és problema en aquest cas, ja que donada la naturalesa de les variables aquest arrodoniment no provoca cap pèrdua d'informació ni ambigüïtat entre valors.

Finalment, com que la base de dades no conté *missing values*, s'afegeixen uns quants artificialment de forma aleatòria per tal de poder treballar les tècniques de *missing values* vistes a classe. Aquests *missing values* simulats es creen proporcionalment i s'ha imposat que suposin un 5% dels valors de la base de dades.

#### 3.2 Anàlisi descriptiva abans del preprocessing

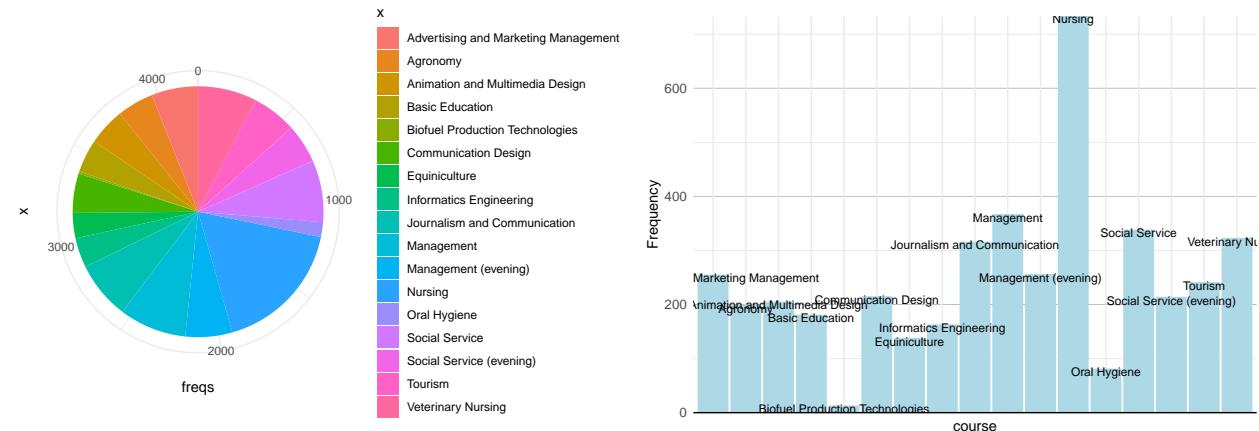
En aquesta secció es fa una anàlisi descriptiva (univariant i bivariant) de les dades crues abans del preprocessing.

##### 3.2.1 Anàlisi univariant de les variables

###### Variables categòriques

Pel que fa a l'anàlisi descriptiva univariant per a variables categòriques, es creen gràfiques de tipus *pie chart* que representen la part proporcional de cada nivell o categoria sobre el total d'observacions de la variable categòrica així com un gràfic de barres que compta les ocurrències de cada categoria o nivell de la variable. La mateixa informació que mostren aquests dos gràfics es pot representar en una taula de freqüències.

*COURSE*



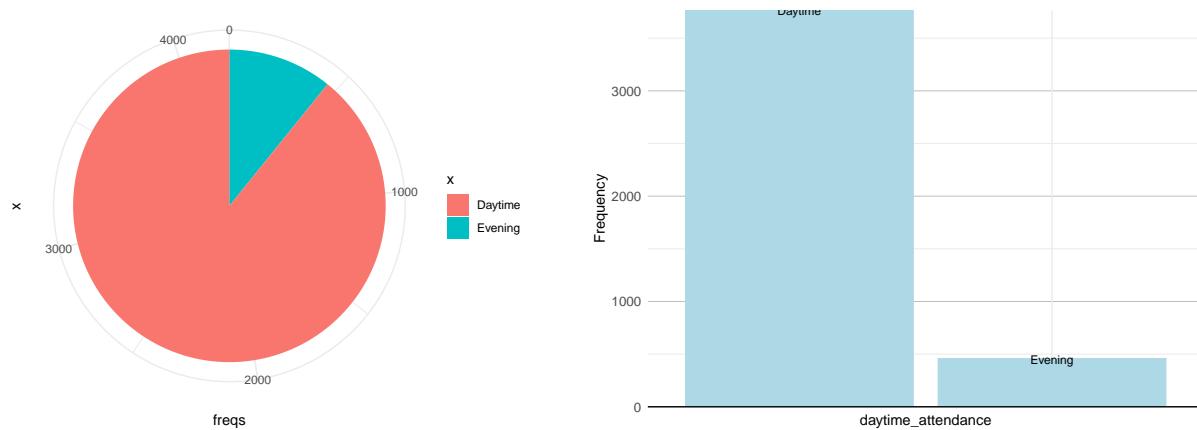
```

## 
## -----
## Variable Stats / Values Freqs (% of Valid) Missing
## -----
## course\ 1\. Biofuel Production Techno\ 12 ( 0.3%)\ 202\
## [factor] 2\. Animation and Multimedia \ 204 ( 4.8%)\ (4.6%)
## 3\. Social Service (evening)\ 213 ( 5.0%)\ 
## 4\. Agronomy\ 198 ( 4.7%)\ 
## 5\. Communication Design\ 214 ( 5.1%)\ 
## 6\. Veterinary Nursing\ 322 ( 7.6%)\ 
## 7\. Informatics Engineering\ 162 ( 3.8%)\ 
## 8\. Equiniculture\ 137 ( 3.2%)\ 
## 9\. Management\ 366 ( 8.7%)\ 
## 10\. Social Service\ 337 ( 8.0%)\ 
## [ 7 others ] 2057 (48.7%)
## -----

```

El curs amb menys estudiants és el de “Biofuel Production Techno” amb un total de 12 alumnes. Així mateix, el curs amb més estudiants és el de “Nursing”.

#### *DAYTIME\_ATTENDANCE*



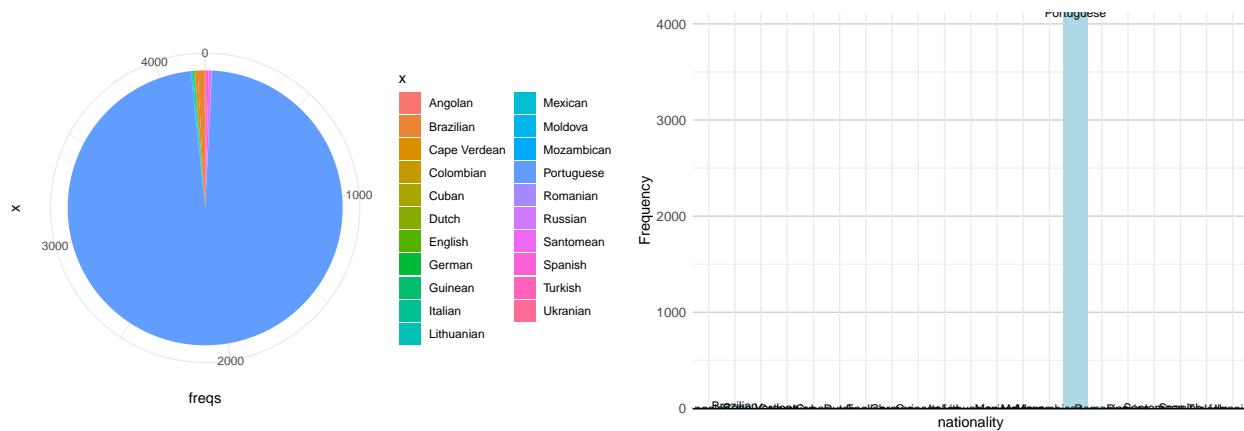
```

## 
## -----
## Variable Stats / Values Freqs (% of Valid) Missing
## -----
## daytime_attendance\ 1\. Evening\ 456 (10.8%)\ 205\
## [factor] 2\. Daytime 3763 (89.2%) (4.6%)
## -----

```

La major part de l'assistència es concentra en la franja horària del matí.

## NATIONALITY

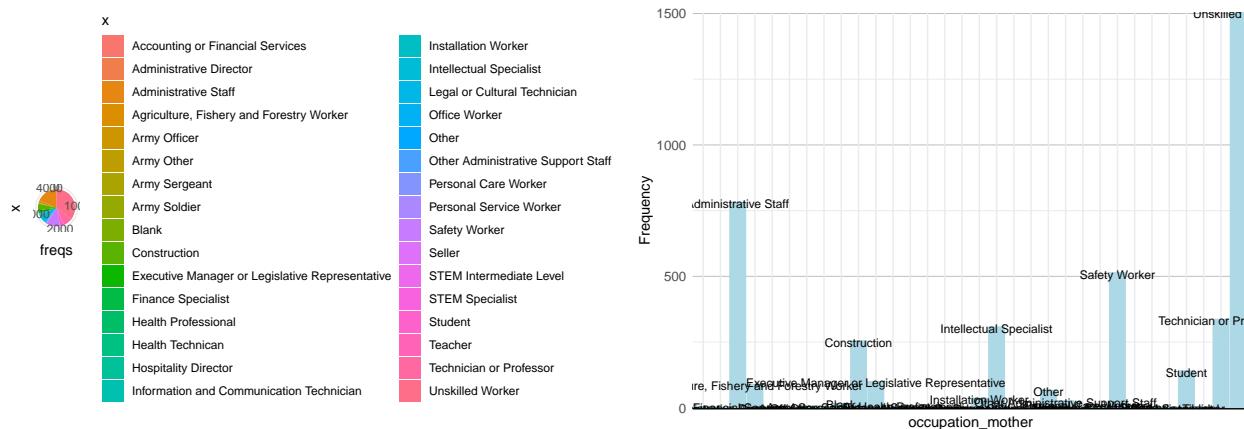


```
##
```

## Variable	Stats / Values	Freqs (% of Valid)	Missing
## nationality\	1\. Portuguese\	4119 (97.5%)\ 200\	
## [factor]	2\. German\	2 ( 0.0%)\ (4.5%)	
##	3\. Spanish\	13 ( 0.3%)\ 1	
##	4\. Italian\	3 ( 0.1%)\ 1	
##	5\. Dutch\	1 ( 0.0%)\ 1	
##	6\. English\	1 ( 0.0%)\ 1	
##	7\. Lithuanian\	1 ( 0.0%)\ 1	
##	8\. Angolan\	2 ( 0.0%)\ 1	
##	9\. Cape Verdean\	13 ( 0.3%)\ 1	
##	10\. Guinean\	5 ( 0.1%)\ 1	
##	[ 11 others ]	64 ( 1.5%)	

El 97.5% dels alumnes són portuguesos. Les altres 20 nacionalitats es distribueixen entre el 2.5% restant dels individus.

## OCCUPATION\_MOTHER

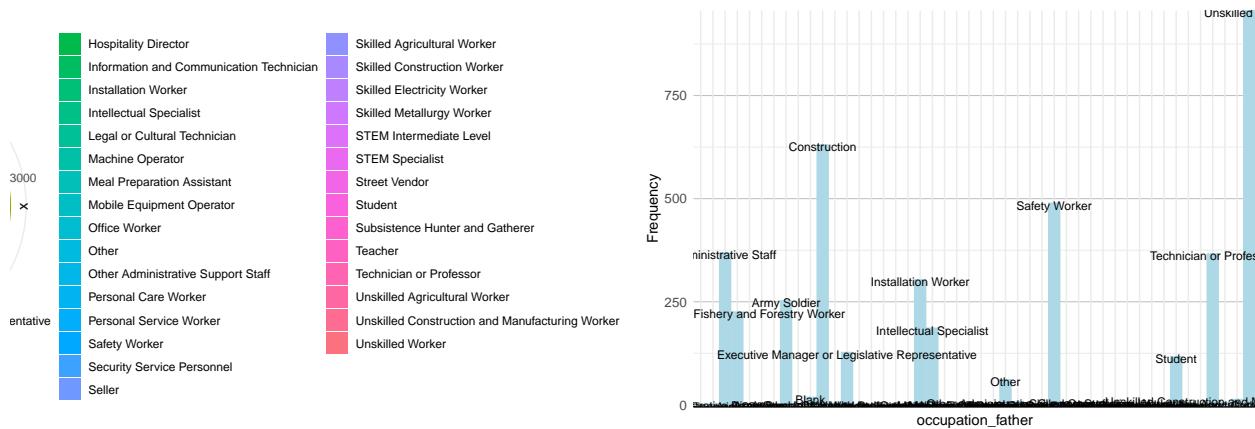


```
##
```

##	## Variable	Stats / Values	Freqs (% of Valid)	Missing
##	## occupation_mother\\	1\\. Student\\	140 ( 3.3%)\\	190\\
## [factor]		2\\. Executive Manager or Legi\\	99 ( 2.3%)\\	(4.3%)
##		3\\. Intellectual Specialist\\	307 ( 7.3%)\\	
##		4\\. Technician or Professor\\	336 ( 7.9%)\\	
##		5\\. Administrative Staff\\	781 (18.4%)\\	
##		6\\. Safety Worker\\	513 (12.1%)\\	
##		7\\. Agriculture, Fishery and \\	88 ( 2.1%)\\	
##		8\\. Construction\\	254 ( 6.0%)\\	
##		9\\. Installation Worker\\	36 ( 0.9%)\\	
##		10\\. Unskilled Worker\\	1504 (35.5%)\\	
##		[ 22 others ]	176 ( 4.2%)	
##				

Les ocupacions de les mares més destacades són treballadora no qualificada (35.5%) i personal administratiu (18.4%).

*OCCUPATION\_FATHER*



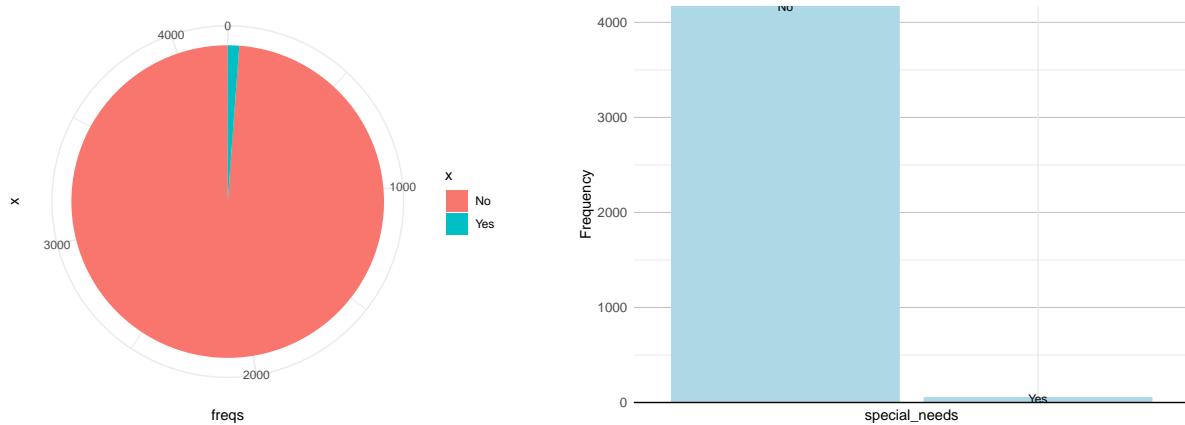
```

## 
## -----
## Variable           Stats / Values          Freqs (% of Valid)  Missing
## -----
## occupation_father\ 1\. Student\           117 ( 2.8%)\
## [factor]           2\. Executive Manager or Legi\ 127 ( 3.0%)\
##                   3\. Intellectual Specialist\ 186 ( 4.4%)\
##                   4\. Technician or Professor\ 366 ( 8.7%)\
##                   5\. Administrative Staff\ 368 ( 8.8%)\
##                   6\. Safety Worker\   489 (11.7%)\
##                   7\. Agriculture, Fishery and \
##                   8\. Construction\ 630 (15.1%)\
##                   9\. Installation Worker\ 303 ( 7.2%)\
##                   10\. Unskilled Worker\ 956 (22.9%)\
##                   [ 36 others ]      416 ( 9.9%)\
## -----

```

La feina més representativa dels pares és treballador no qualificat, seguida dels llocs de treball a la construcció.

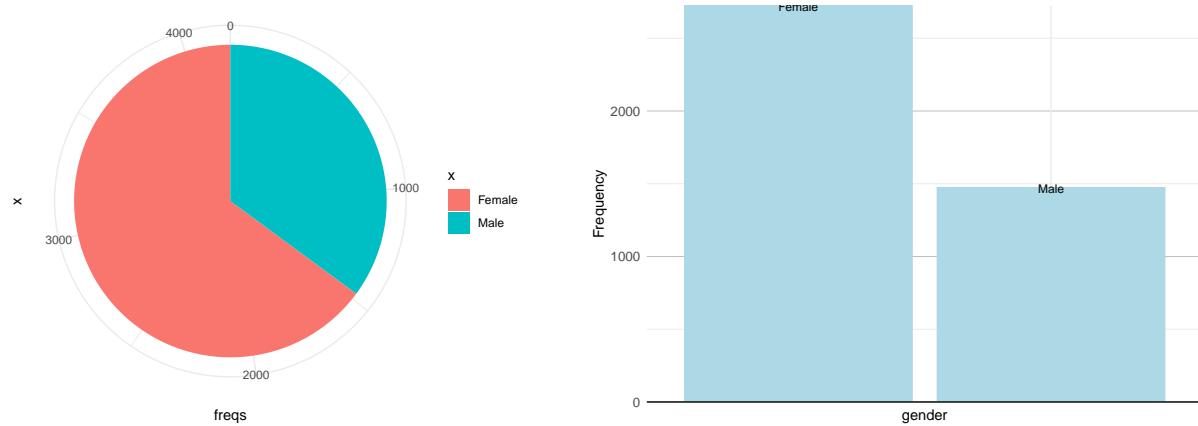
## SPECIAL\_NEEDS



```
##  
## -----  
## Variable      Stats / Values   Freqs (% of Valid)  Missing  
## -----  
## special_needs\ 1\. No\        4170 (98.8%)\    204\  
## [factor]       2\. Yes         50 ( 1.2%)       (4.6%)  
## -----
```

El 98.9% dels alumnes no tenen necessitats especials.

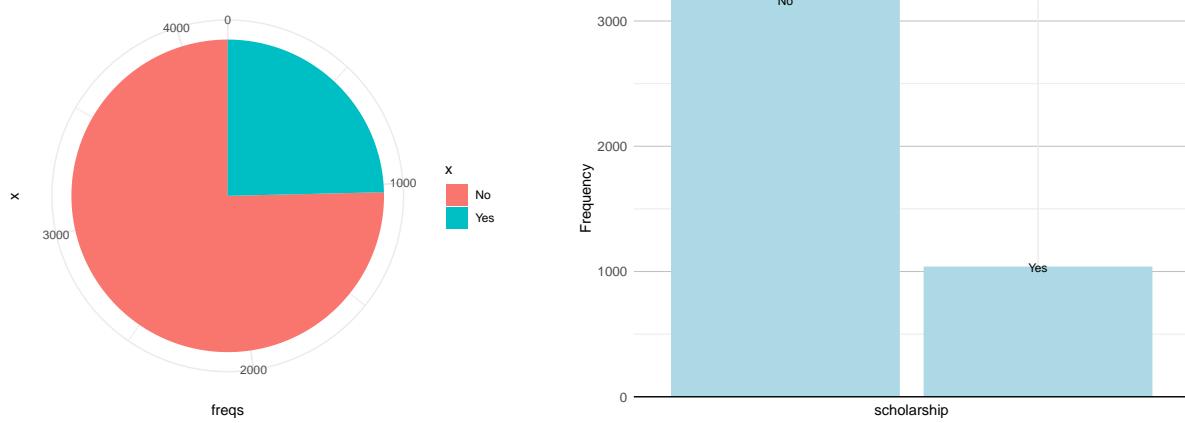
## GENDER



```
##  
## -----  
## Variable      Stats / Values   Freqs (% of Valid)  Missing  
## -----  
## gender\       1\. Female\     2724 (64.9%)\    227\  
## [factor]       2\. Male        1473 (35.1%)      (5.1%)  
## -----
```

Aproximadament el 65% dels individus s'identifiquen com a dona.

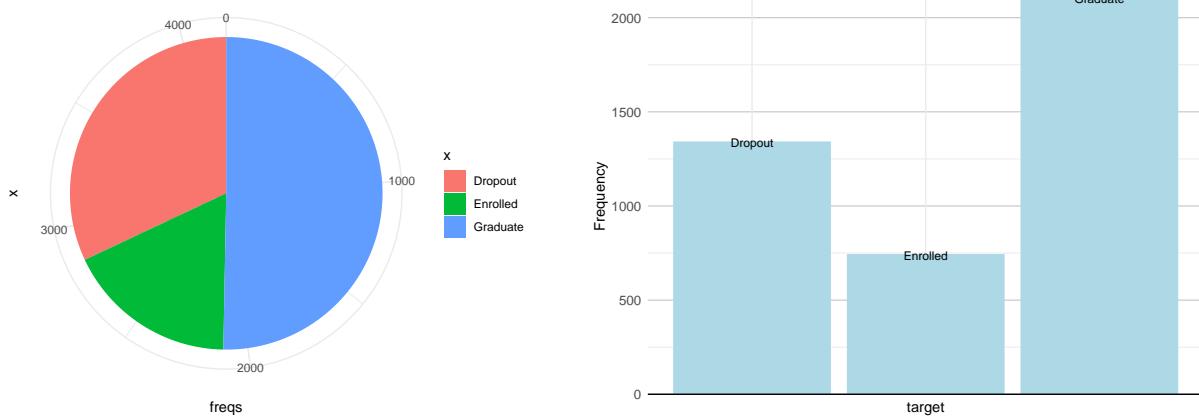
## SCHOLARSHIP



```
##  
## -----  
## Variable      Stats / Values   Freqs (% of Valid)  Missing  
## -----  
## scholarship\  1\. No\        3164 (75.4%)\  
## [factor]       2\. Yes        1034 (24.6%)  
## (5.1%)  
## -----
```

Tres quartes parts dels alumnes no reben cap mena d'ajuda econòmica.

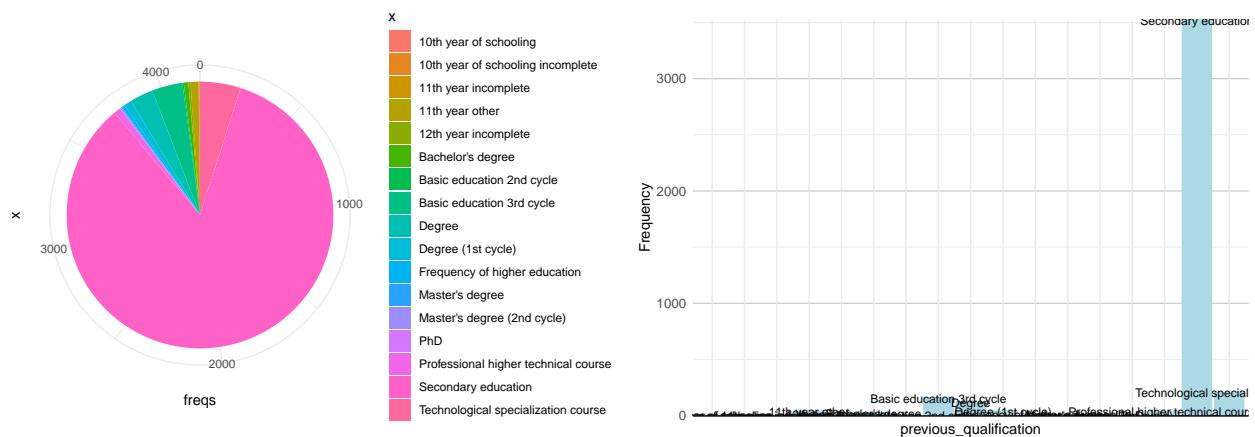
## TARGET



```
##  
## -----  
## Variable      Stats / Values   Freqs (% of Valid)  Missing  
## -----  
## target\       1\. Dropout\    1339 (32.0%)\  
## [factor]       2\. Enrolled\   741 (17.7%)  
## (5.4%)  
## 3\. Graduate   2105 (50.3%)  
## -----
```

La meitat dels enquestats són graduats, el 32% ha abandonat els estudis i quasi el 18% són personnes que han reprès els estudis.

*PREVIOUS\_QUALIFICATION*



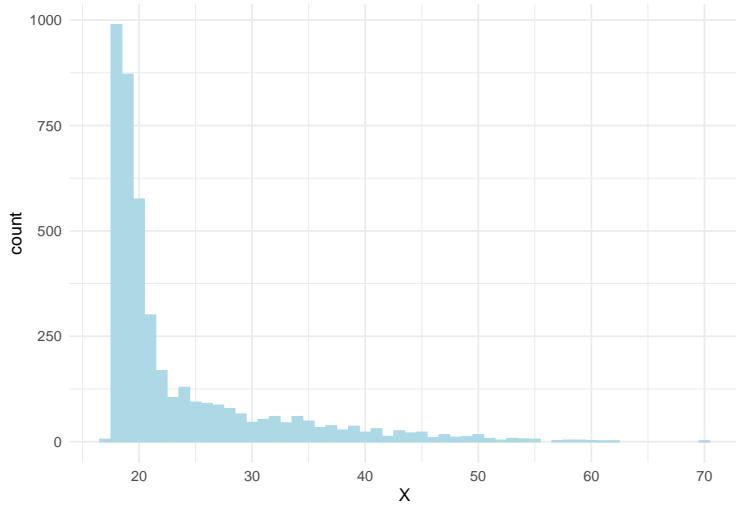
##	##	##	##	##
## Variable	Stats / Values	Freqs (% of Valid)	##	Missing
## previous_qualification\\ [factor]	1\\. Secondary education\\ 2\\. Bachelor's degree\\ 3\\. Degree\\ 4\\. Master's degree\\ 5\\. PhD\\ 6\\. Frequency of higher educa\\ 7\\. 12th year incomplete\\ 8\\. 11th year incomplete\\ 9\\. 11th year other\\ 10\\. 10th year of schooling\\ [ 7 others ]	3526 (84.0%)\\ 23 ( 0.5%)\\ 119 ( 2.8%)\\ 6 ( 0.1%)\\ 1 ( 0.0%)\\ 15 ( 0.4%)\\ 11 ( 0.3%)\\ 4 ( 0.1%)\\ 41 ( 1.0%)\\ 1 ( 0.0%)\\ 452 (10.8%)	##	225\\ (5.1%)

El 84% dels estudiants tenen l'Educació Secundària Obligatòria.

## Variables numériques

Per a l'anàlisi univariant de variables numèriques s'utilitza un histograma per analitzar visualment els valors que pren aquesta variable. També es fa un resum (i.e. *summary*) de la variable que permet conèixer, entre altres coses, quina és la mitjana, la mediana, el valor màxim i el valor mínim que pren.

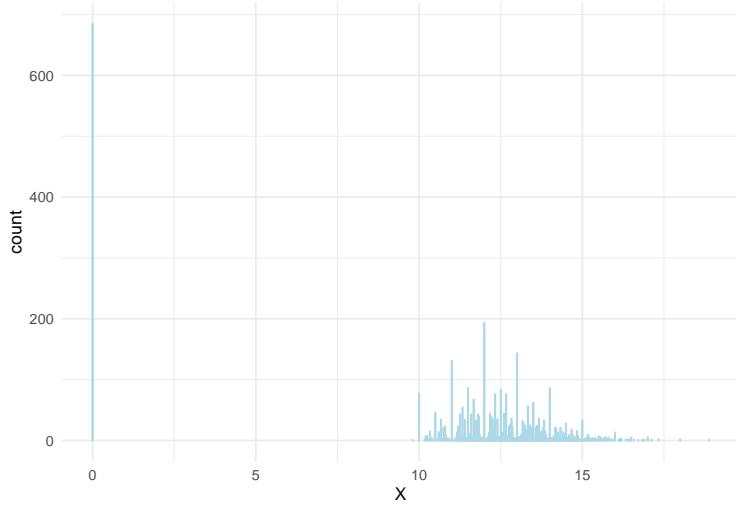
### *AGE*



```
##  
## | N.Valid | Min | Q1 | Median | Mean | Std.Dev | Q3 | Max | IQR |  
## |:-----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|  
## | 4226.000 | 17.000 | 19.000 | 20.000 | 23.245 | 7.555 | 25.000 | 70.000 | 6.000 |
```

La variable edat es comprèn dins un interval que va dels 17 als 70 anys amb ambdós inclosos. No obstant això, la població es distribueix principalment al voltant dels 20 anys.

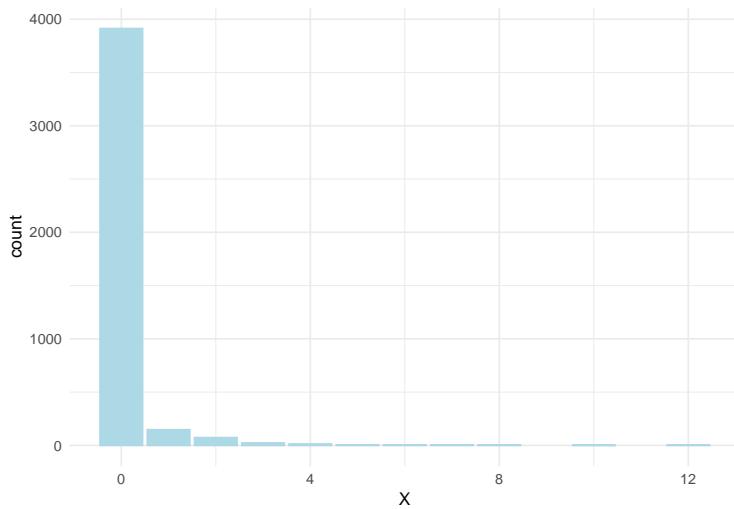
### *GRADES\_SEM\_1*



```
##  
## | N.Valid | Min | Q1 | Median | Mean | Std.Dev | Q3 | Max | IQR |  
## |:-----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|  
## | 4225.000 | 0.000 | 11.000 | 12.290 | 10.645 | 4.842 | 13.400 | 18.880 | 2.400 |
```

Les notes del primer semestre es reparteix en gran manera entre 10 i 14.

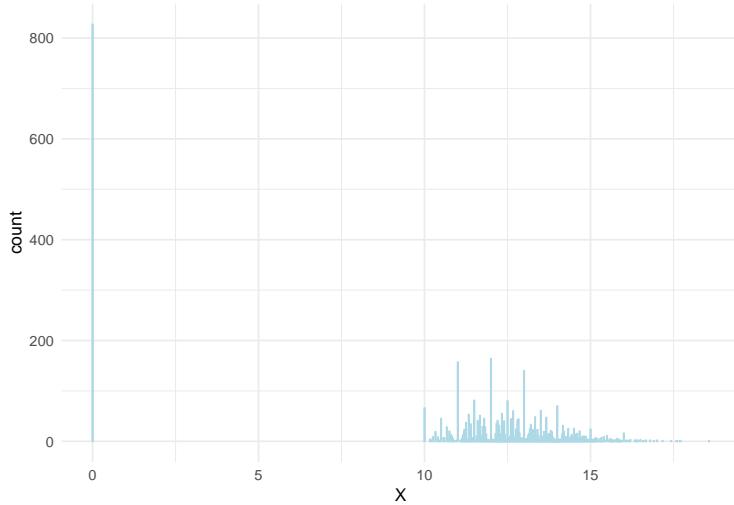
### *NO\_EVAL\_SEM\_1*



```
##  
## | N.Valid | Min | Q1 | Median | Mean | Std.Dev | Q3 | Max | IQR |  
## |:-----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|  
## | 4191.000 | 0.000 | 0.000 | 0.000 | 0.136 | 0.686 | 0.000 | 12.000 | 0.000 |
```

Majoritàriament, no hi ha unitats curriculars no avaluades el primer semestre.

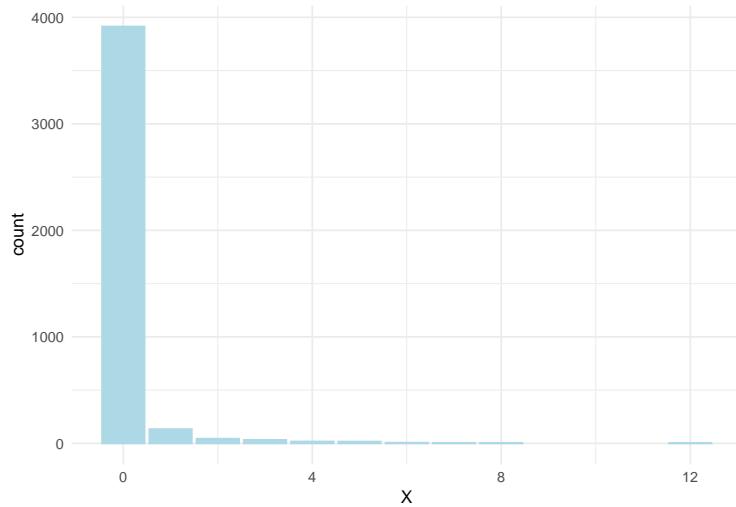
### *GRADES\_SEM\_2*



```
##  
## | N.Valid | Min | Q1 | Median | Mean | Std.Dev | Q3 | Max | IQR |  
## |:-----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|  
## | 4179.000 | 0.000 | 10.750 | 12.210 | 10.219 | 5.223 | 13.330 | 18.570 | 2.580 |
```

Les notes del segon semestre es reparteix en gran manera entre 10 i 14.

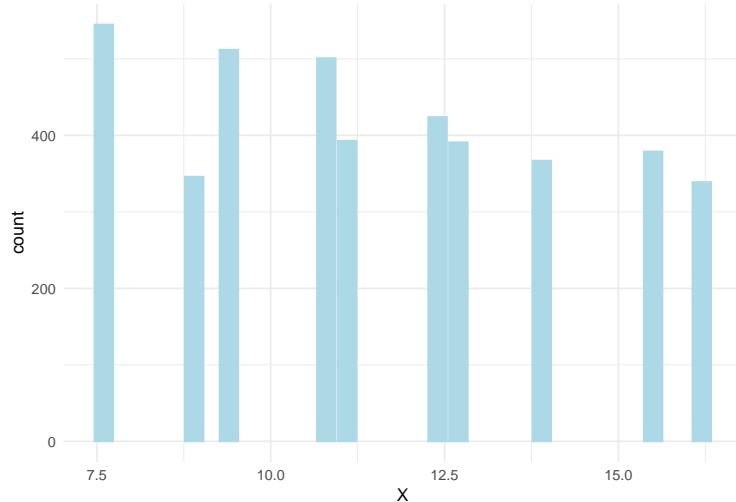
### *NO\_EVAL\_SEM\_2*



```
##  
## | N.Valid | Min | Q1 | Median | Mean | Std.Dev | Q3 | Max | IQR |  
## |:-----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|  
## | 4179.000 | 0.000 | 0.000 | 0.000 | 0.148 | 0.747 | 0.000 | 12.000 | 0.000 |
```

Majoritàriament, no hi ha unitats no avaluades el segon semestre.

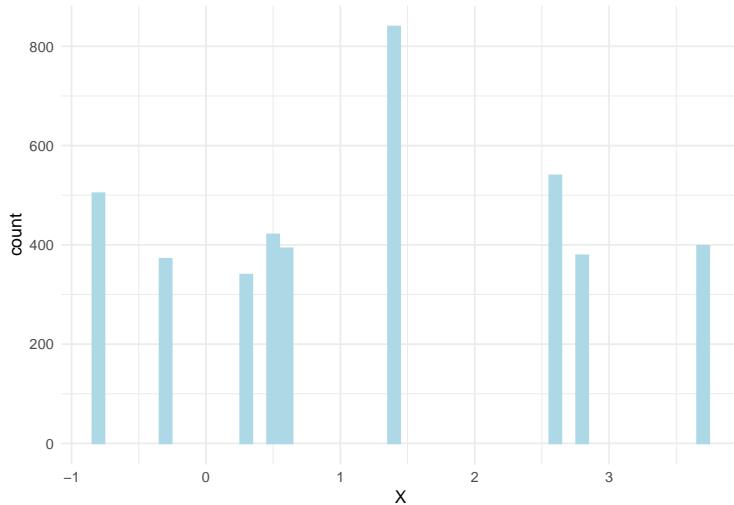
### *UNEMPLOYMENT\_RATE*



```
##  
## | N.Valid | Min | Q1 | Median | Mean | Std.Dev | Q3 | Max | IQR |  
## |:-----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|  
## | 4197.000 | 7.600 | 9.400 | 11.100 | 11.555 | 2.663 | 13.900 | 16.200 | 4.500 |
```

El valor més baix que pren la taxa d'atur és 7.6% i el més alt, 16.2%.

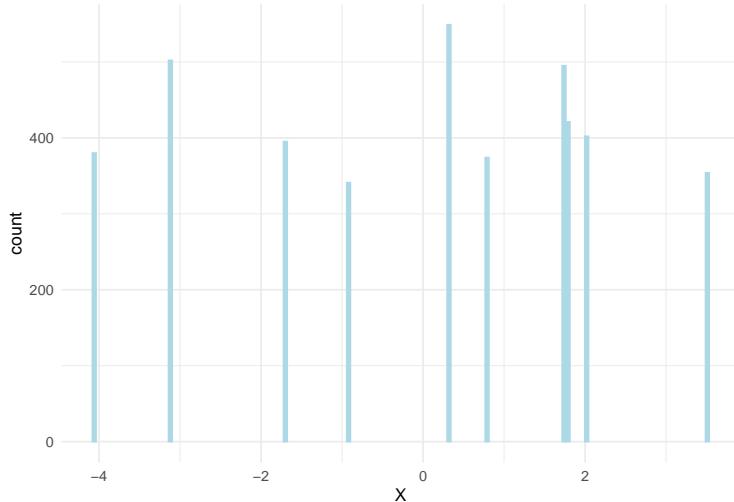
### *INFLATION\_RATE*



```
##
## | N.Valid | Min | Q1 | Median | Mean | Std.Dev | Q3 | Max | IQR |
## |:-----:|:----:|:---:|:---:|:---:|:---:|:---:|:---:|:---:|:---:|
## | 4187.000 | -0.800 | 0.300 | 1.400 | 1.229 | 1.384 | 2.600 | 3.700 | 2.300 |
```

La màxima taxa d'inflació és del 3.7%.

### *GDP*



```
##
## | N.Valid | Min | Q1 | Median | Mean | Std.Dev | Q3 | Max | IQR |
## |:-----:|:----:|:---:|:---:|:---:|:---:|:---:|:---:|:---:|:---:|
## | 4213.000 | -4.060 | -1.700 | 0.320 | 0.011 | 2.271 | 1.790 | 3.510 | 3.490 |
```

El PIB oscil·la entre -4 i 3.5.

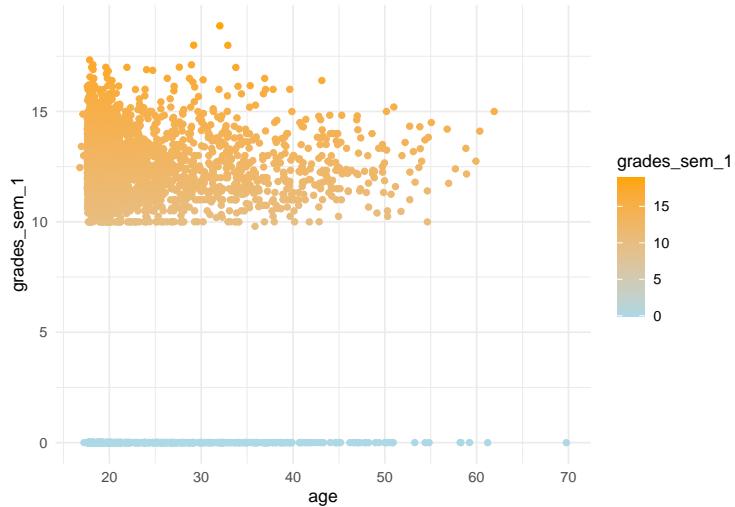
### 3.2.2 Anàlisi bivariant

Per a l'anàlisi bivariant es miren les combinacions més interessants (i que més sentit té fer) de dues variables.

#### Dues variables numèriques

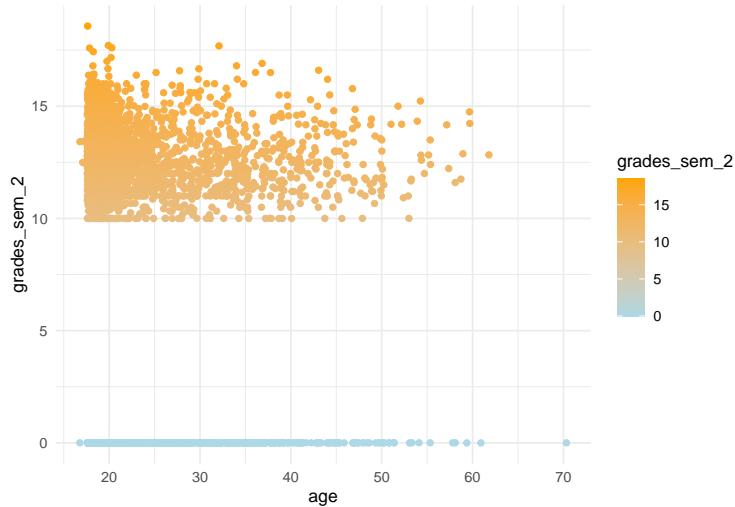
A l'hora d'analitzar conjuntament dues variables numèriques, s'utilitza un *scatterplot* on cada variable està representada en un dels dos eixos.

*AGE* i *GRADES\_SEM\_1*



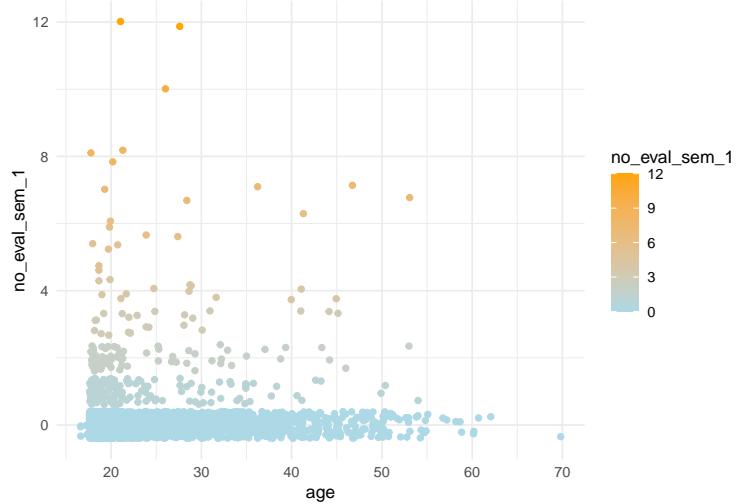
Es veu que no hi ha diferències significatives en la nota del primer semestre segons l'edat. Però en les edats més avançades, les notes poden no arribar a ser tan altes.

*AGE* i *GRADES\_SEM\_2*



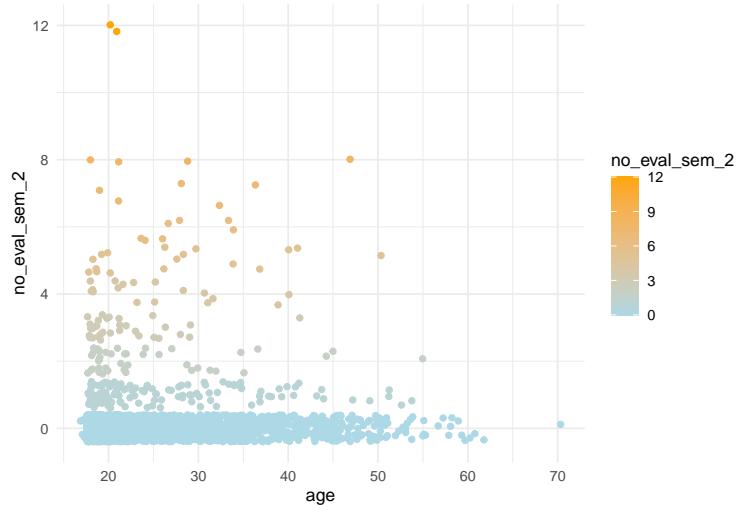
S'observa el mateix que en el gràfic anterior amb les notes del primer semestre.

### AGE i NO\_EVAL\_SEM\_1



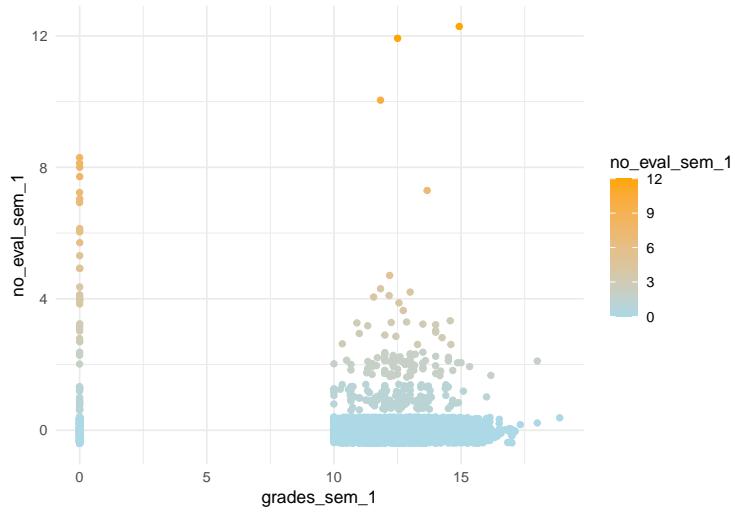
Segons l'edat, no sembla que hi hagi diferències en les unitats no evaluades.

### AGE i NO\_EVAL\_SEM\_2



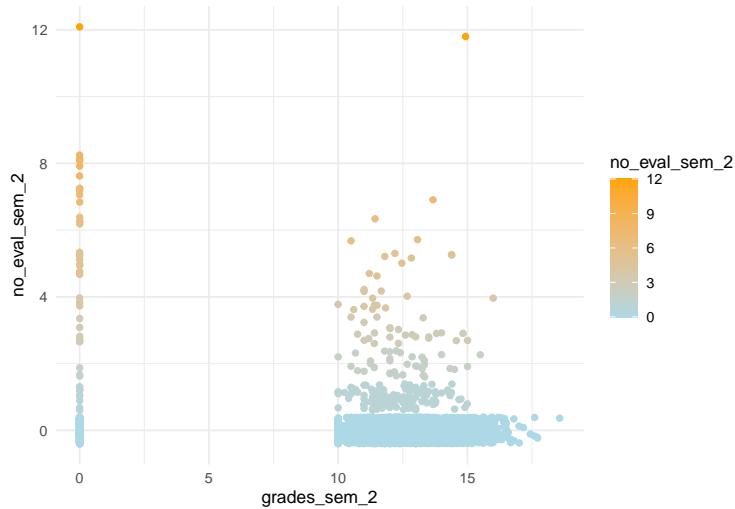
Ocorre el mateix que en les unitats no evaluades del primer semestre, l'edat no sembla ser significativa.

### *GRADES\_SEM\_1* i *NO\_EVAL\_SEM\_1*



Si es mira el gràfic, no s'observa gaire correlació entre les notes i les unitats no avaluades del primer semestre.

### *GRADES\_SEM\_2* i *NO\_EVAL\_SEM\_2*



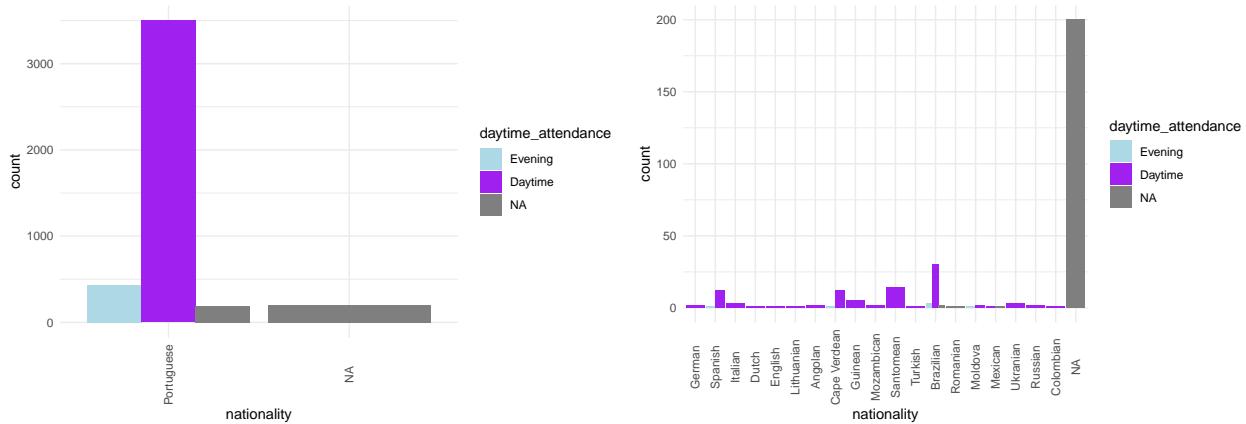
No s'observa gaire correlació entre les notes i les unitats no avaluades del segon semestre.

### Dues variables categòriques

Quan hi ha dues variables categòriques, s'utilitza un gràfic de barres per representar les ocurrències d'una de les variables separant entre nivells o categories de l'altra variable.

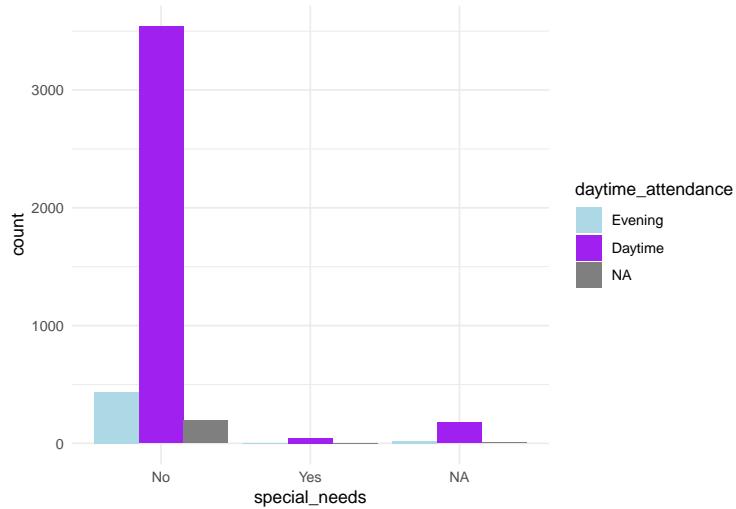
### *NATIONALITY* i *DAYTIME\_ATTENDANCE*

Per a aquesta primera anàlisi, en ser la majoria dels alumnes de nacionalitat portuguesa se separen el nombre de casos de persones portugueses respecte als altres casos, ja que si no, no es pot apreciar bé en el gràfic com es distribueixen els casos dels individus no portuguesos.



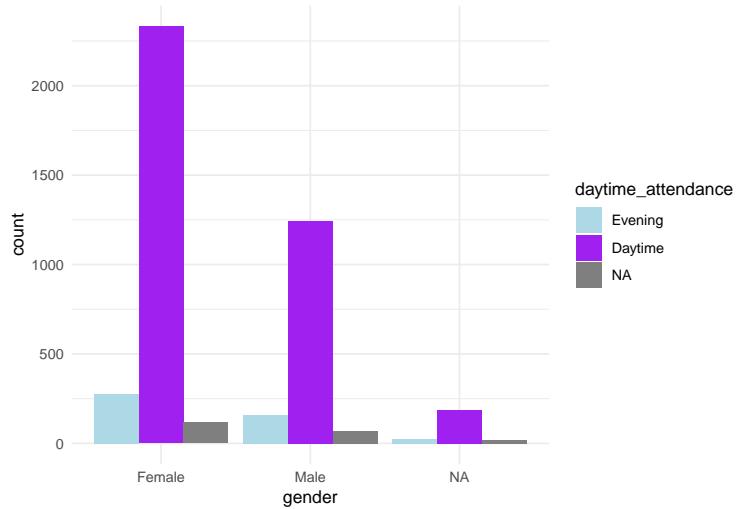
Pel cas de persones portugueses, hi ha major assistència durant el matí. Passa el mateix amb els casos de les altres nacionalitats.

#### *SPECIAL\_NEEDS i DAYTIME\_ATTENDANCE*



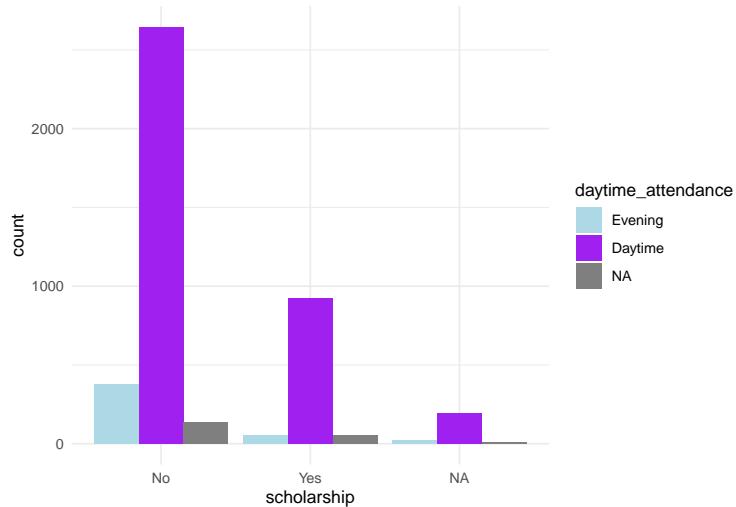
Els individus que no tenen necessitats especials assisteixen majoritàriament al matí, mentre que els que sí que tenen necessitats especials, només assisteixen al matí.

### *GENDER i DAYTIME\_ATTENDANCE*



Independentment del gènere, la majoria va a classe en horari de matí.

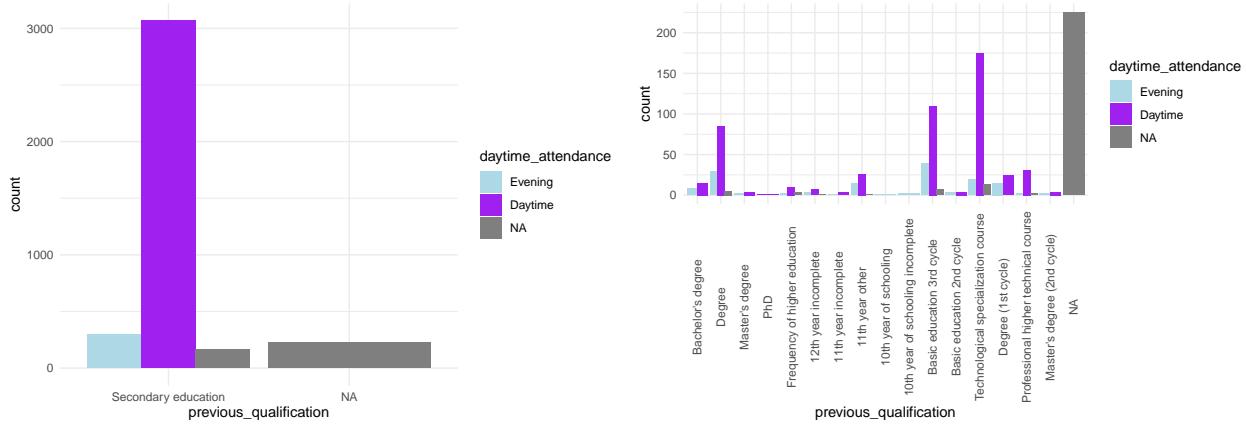
### *SCHOLARSHIP i DAYTIME\_ATTENDANCE*



La majoria de les persones van de matins, sense importar si tenen beca o no.

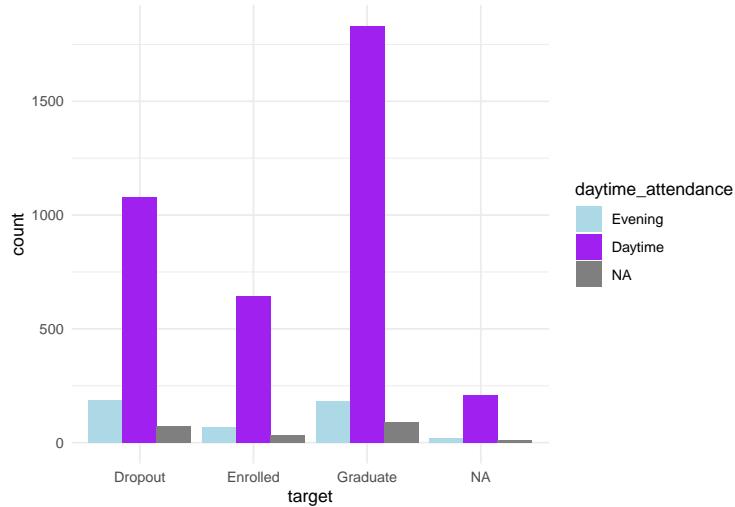
### *PREVIOUS\_QUALIFICATION i DAYTIME\_ATTENDANCE*

Com que la major part dels individus tenen "Secondary education" com a qualificació prèvia, s'han separat aquests casos respecte als altres per tal de poder observar amb millor precisió com es distribueixen.



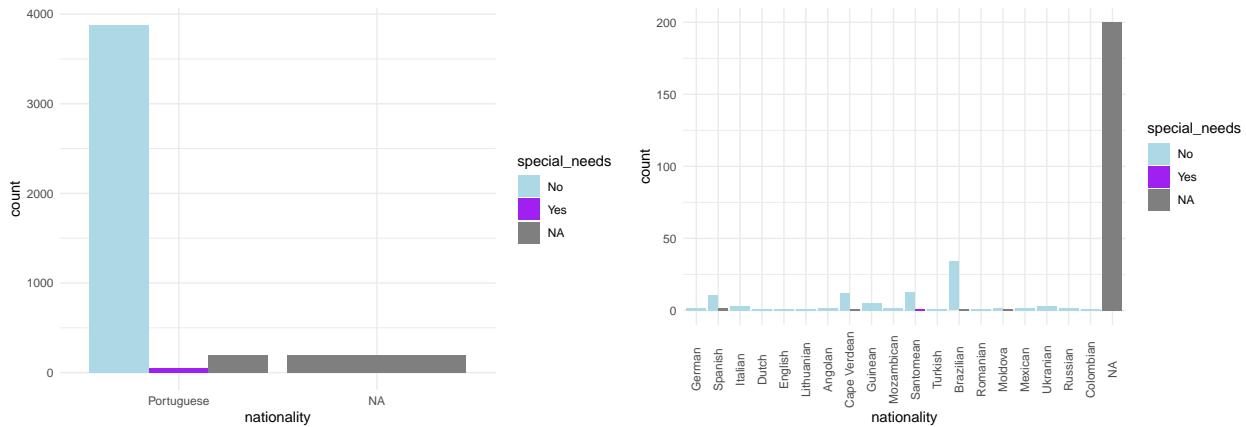
Sembla que independentment de la qualificació prèvia, els individus assisteixen durant el matí.

#### TARGET i DAYTIME\_ATTENDANCE



Independentment de si van acabar els estudis, els van abandonar o els van continuar més tard, hi ha una tendència a anar al matí a classes.

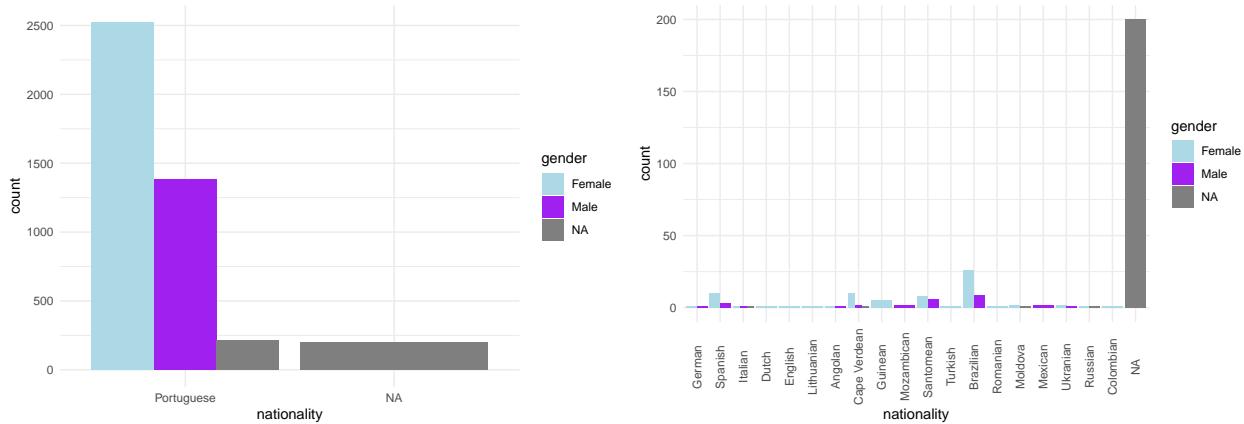
#### NATIONALITY i SPECIAL\_NEEDS



Es pot observar que la majoria d'estudiants de nacionalitat portuguesa no tenen necessitats especials. Per a les altres nacionalitats, no es pot saber significativament la probabilitat de tenir necessitats especials, ja que

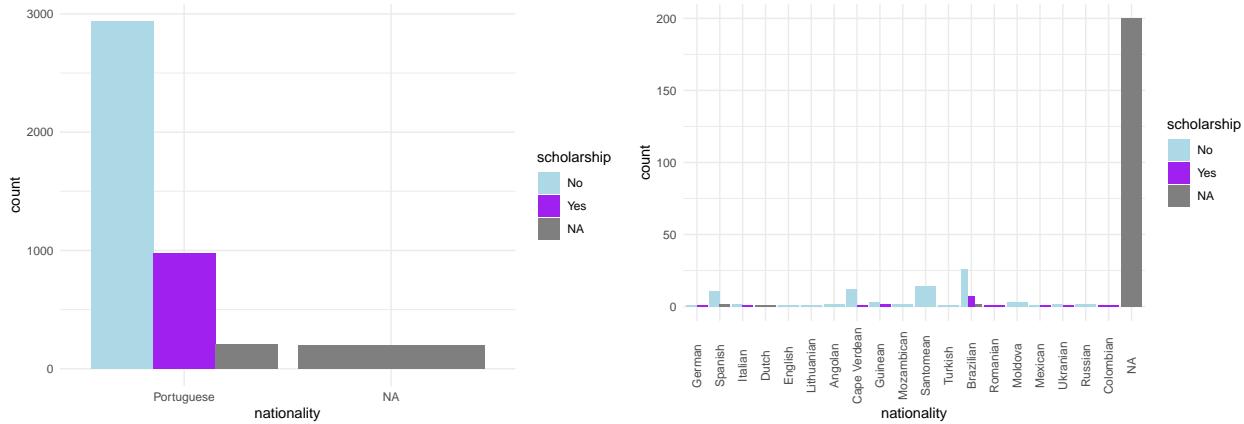
la mida de la mostra és molt petita, però sembla que no tenen necessitats especials.

### NATIONALITY i GENDER



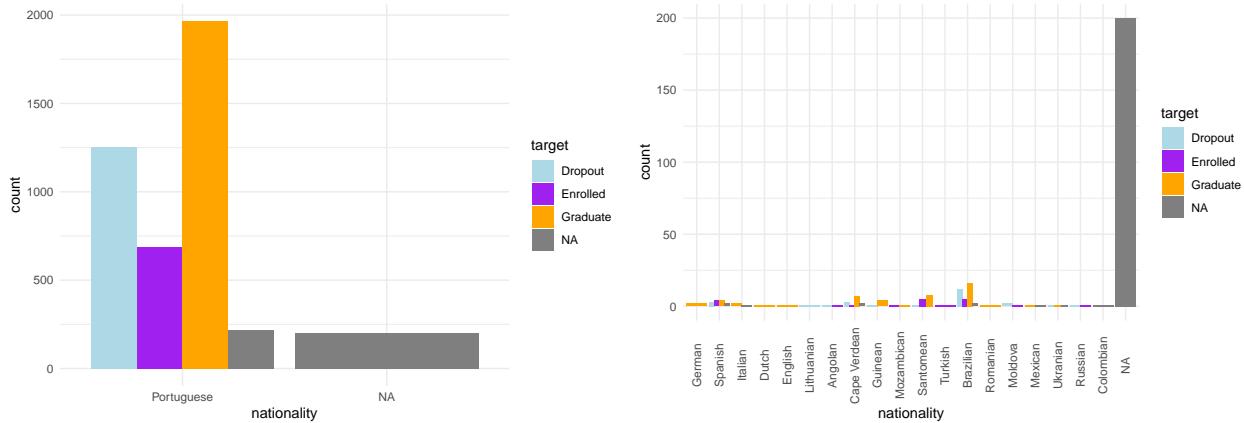
S'observa més tendència a ser estudiant dona que home tant en la nacionalitat portuguesa com en les altres.

### NATIONALITY i SCHOLARSHIP



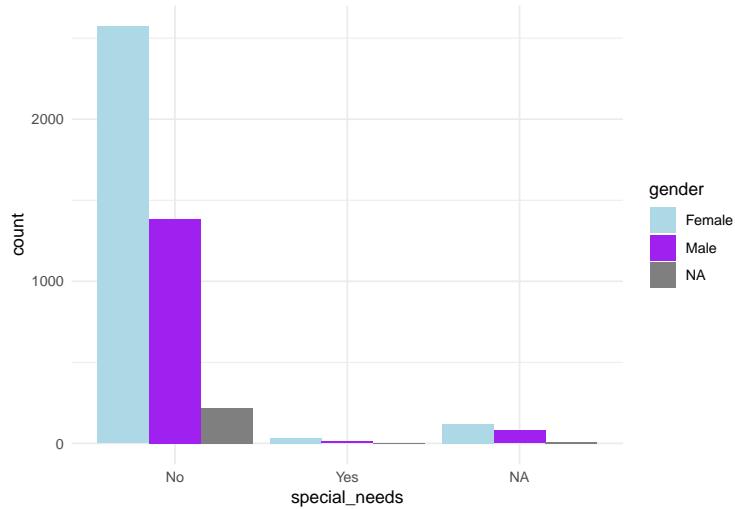
Per a les persones portugueses, la probabilitat de tenir beca és aproximadament un quart. Pels casos de nacionalitat desconeguda s'observa una tendència similar.

### NATIONALITY i TARGET



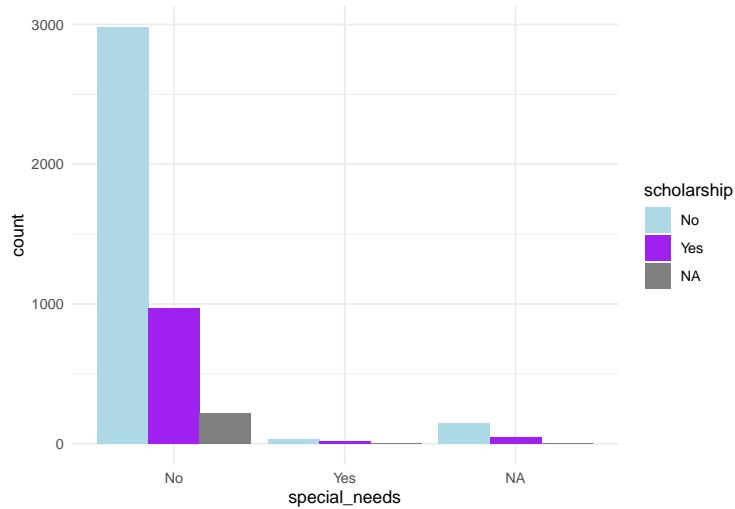
La probabilitat de graduar-se en persones portugueses és més gran que la de deixar la carrera. Ocorre una tendència similar amb les altres nacionalitats.

*SPECIAL\_NEEDS* i *GENDER*



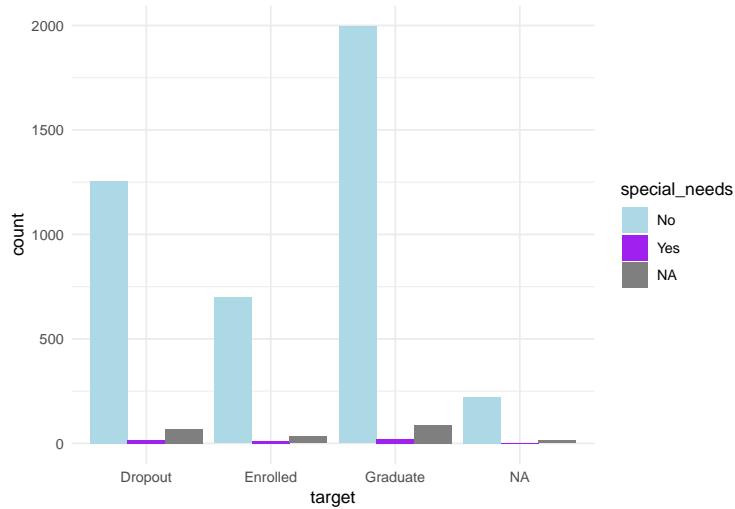
El gènere no sembla rellevant respecte de tenir necessitats especials o no.

*SPECIAL\_NEEDS* i *SCHOLARSHIP*



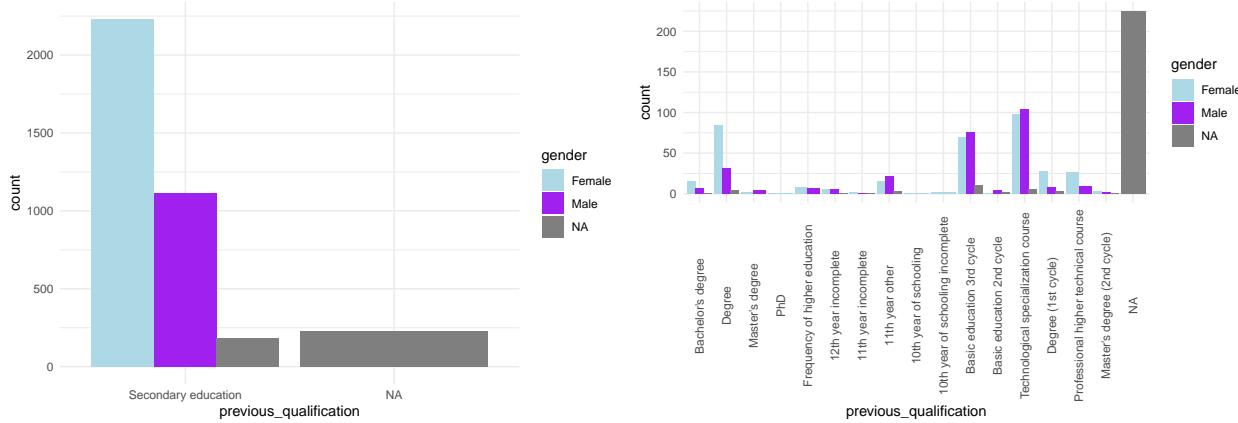
Sembla que en tots els casos, tenint o no necessitats especials, la probabilitat de no rebre una beca és més gran que la probabilitat que la concedeixin.

### TARGET i SPECIAL\_NEEDS



Per a les persones amb necessitats especials s'observa la mateixa tendència que sense necessitats especials, és més probable graduar-se que deixar la carrera.

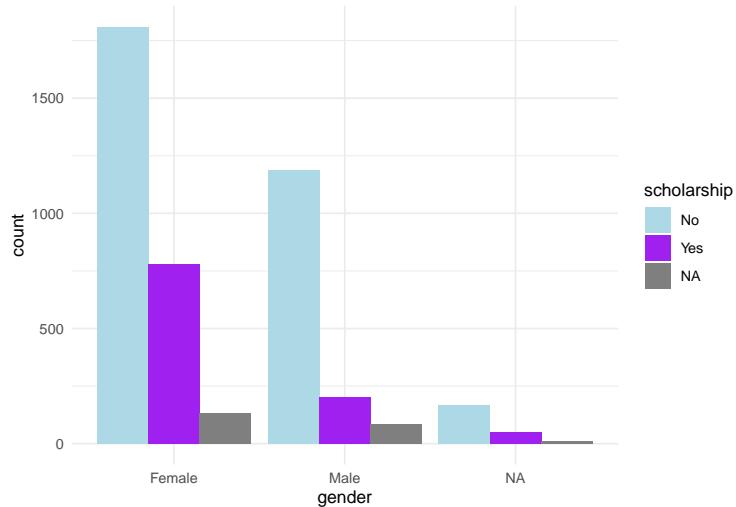
### PREVIOUS\_QUALIFICATION i GENDER



Pel que fa a les persones amb educació secundària com a qualificació prèvia, dos terços són dones.

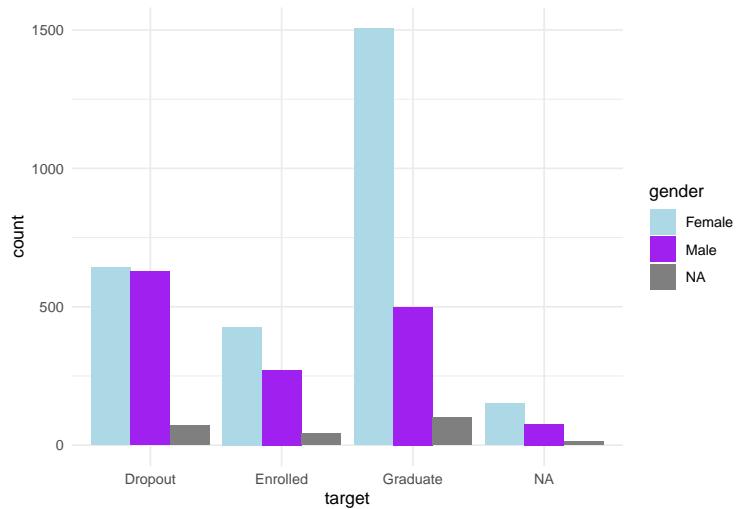
Si s'observen com es distribueix el gènere en les altres qualificacions prèvies, no ocorre el mateix que per a l'educació secundària. Per exemple, hi ha més persones amb un curs d'especialització tecnològica que són homes que dones. No obstant això, per a aquests casos la mida mostra no és gaire gran.

## *GENDER i SCHOLARSHIP*



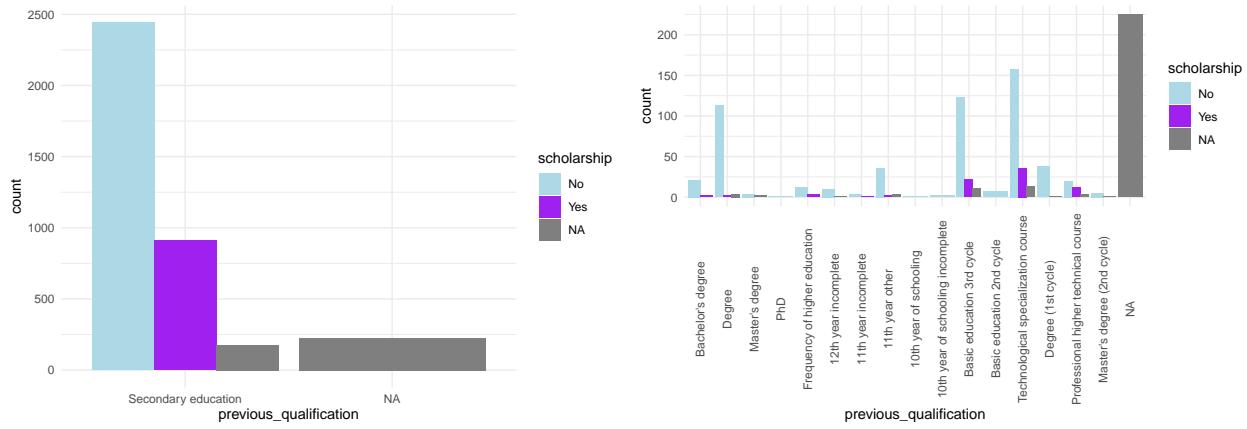
És més probable no tenir “scholarship” que tenir-ne, tant en el cas dels homes com en el de les dones. Tot i això, és més probable tenir beca en el cas de les dones.

## *TARGET i GENDER*



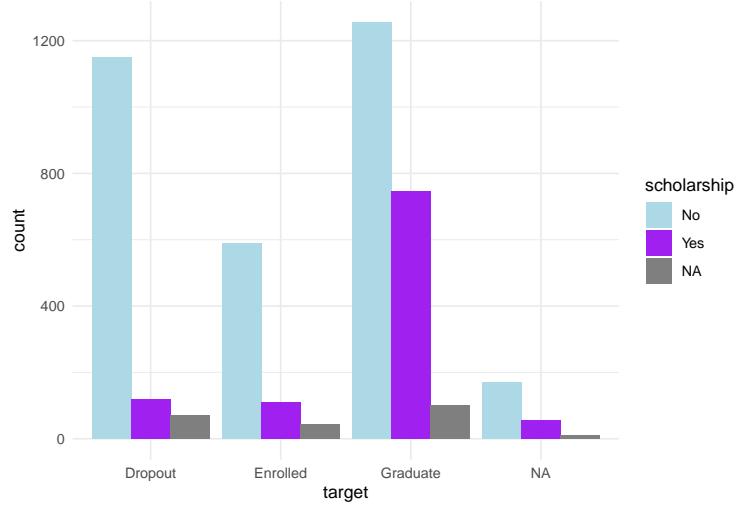
Tenint en compte que hi ha més dones que homes inscrits, la probabilitat de deixar la carrera és més alta com a home.

### *PREVIOUS\_QUALIFICATION i SCHOLARSHIP*



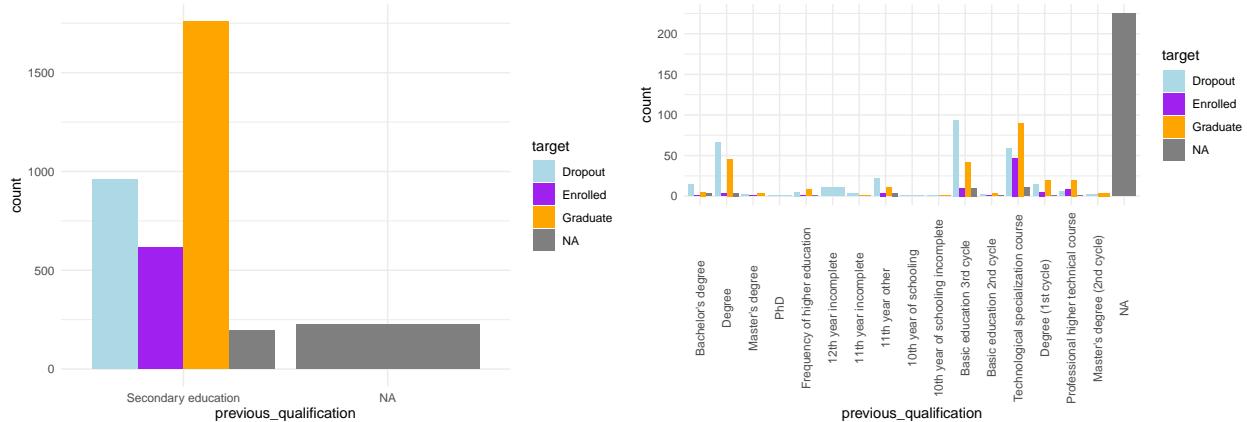
Per a la majoria de les qualificacions prèvies, la probabilitat de tenir beca sembla inferior que la de no tenir-ne.

### *TARGET i SCHOLARSHIP*



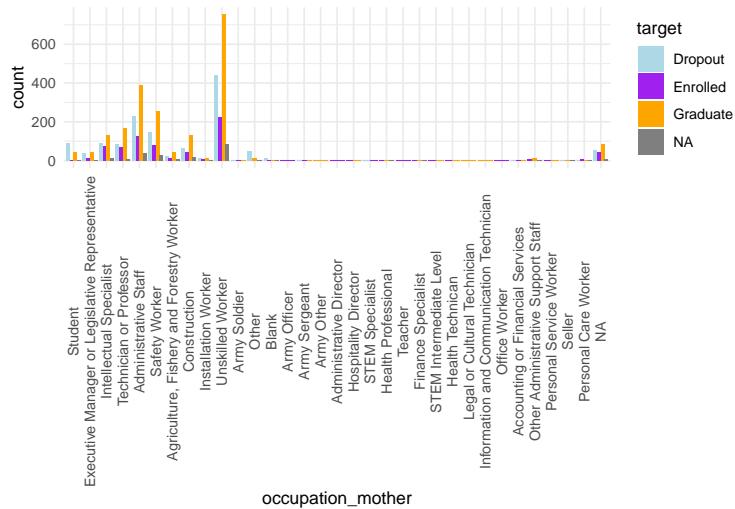
La majoria de les persones amb beca s'han graduat. Les persones sense beca tenen una probabilitat similar de deixar-ho que de graduar-se, aquesta última sent una mica més alta.

### *PREVIOUS\_QUALIFICATION i TARGET*



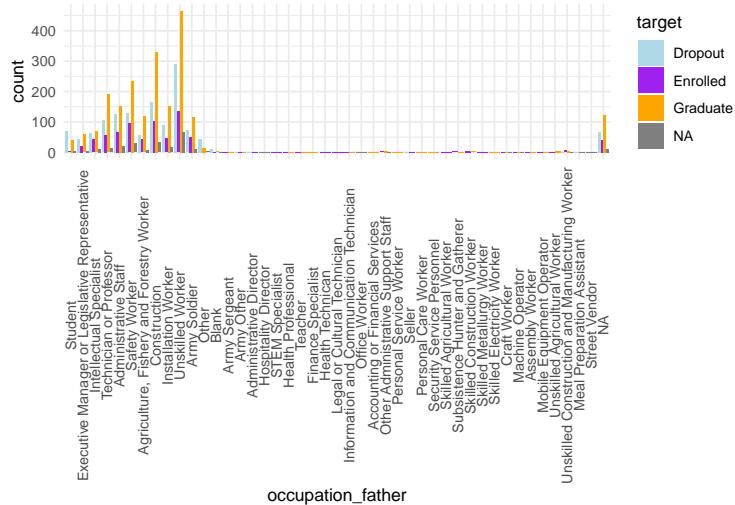
Pels individus que tenen com a qualificació prèvia “Secondary education” tenen més probabilitats de graduarse. No obstant això, per als altres casos la distribució canvia i, per exemple, si tenen “Basic education 3rd cycle” hi ha aproximadament el doble de probabilitats d’abandonar la carrera que de graduar-se.

#### OCCUPATION\_MOTHER i TARGET



L’ocupació de la mare no sembla una variable rellevant per a si la persona es gradua o no.

#### OCCUPATION\_FATHER i TARGET

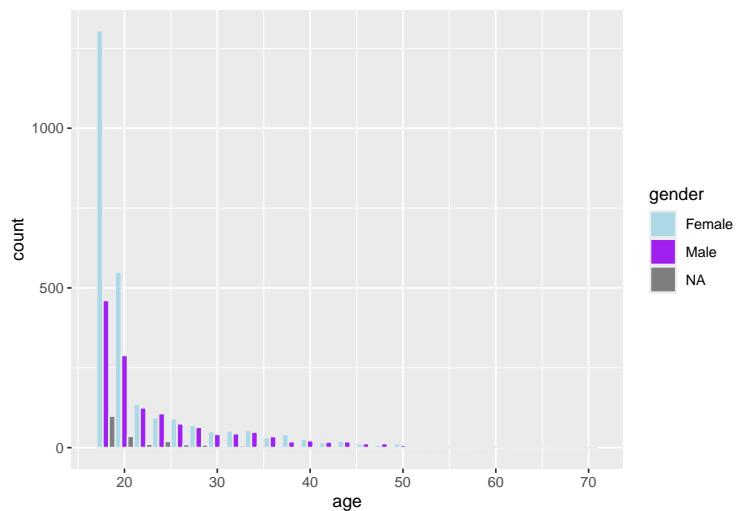


L’ocupació del pare no sembla una variable rellevant per a si la persona es gradua o no.

#### Una variable numèrica i una de categòrica

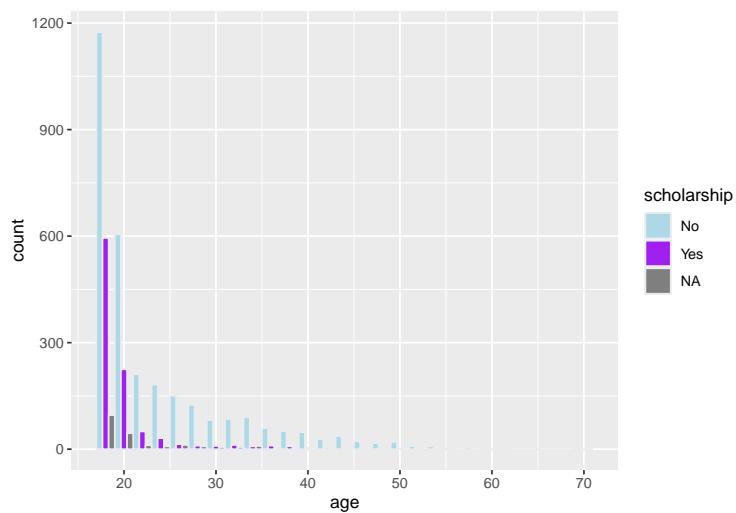
Quan hi ha una variable numèrica i una variable categòrica es fa l’anàlisi multivariant de la mateixa manera que en el cas anterior. La variable numèrica és representada en l’eix de les X i per cada valor d’aquesta variable es representa en format gràfic de barres les ocurrències de cada categoria o nivell de la variable categòrica.

## *AGE i GENDER*



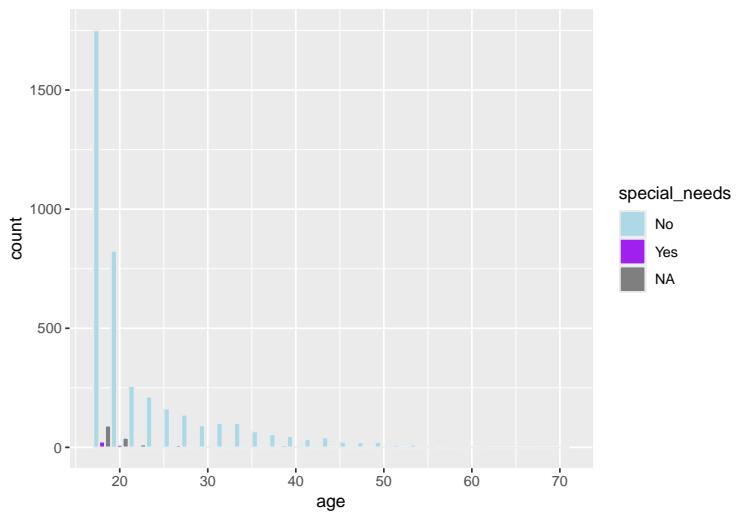
Hi ha més tendència a estudiar en edats joves i en dones. A partir dels 25 anys aproximadament, hi ha un percentatge similar de dones que d'homes que estudien.

## *AGE i SCHOLARSHIP*



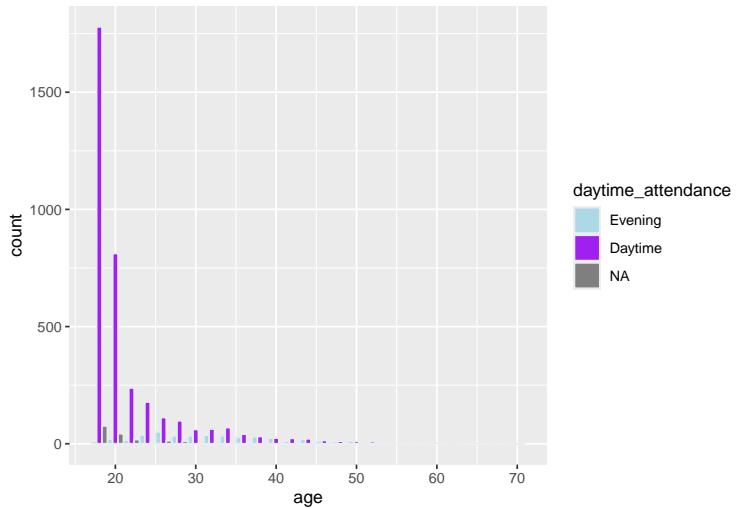
Les beques es concedeixen en edats més joves.

### *AGE i SPECIAL\_NEEDS*



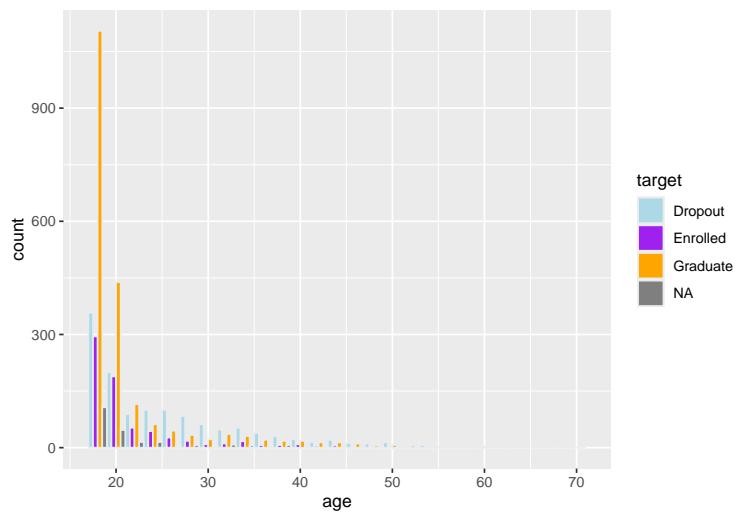
Pel que fa a les necessitats especials, no sembla que estigui correlacionat amb l'edat.

### *AGE i DAYTIME\_ATTENDANCE*



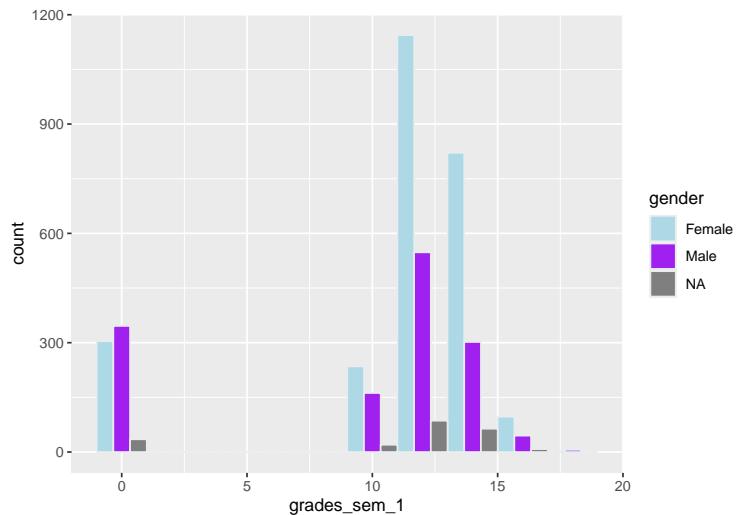
Amb l'edat, incrementa el percentatge d'assistència a les tardes. No obstant això, continua dominant l'assistència al matí en tots els grups d'edats.

### AGE i TARGET



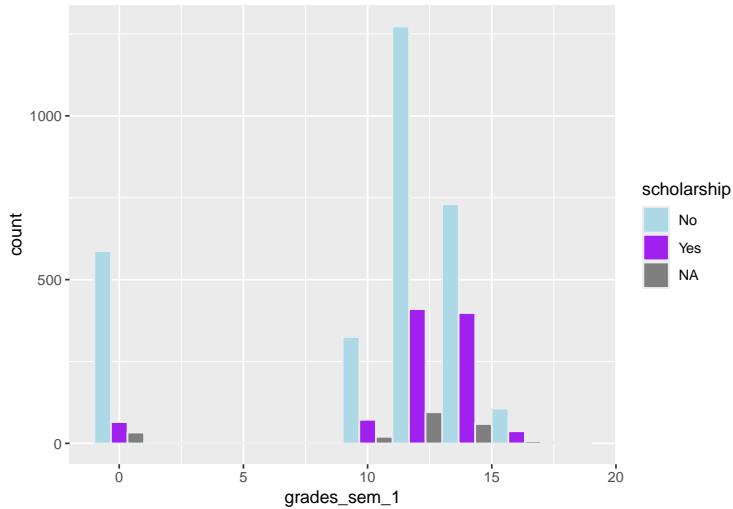
A mesura que avança l'edat, incrementa la probabilitat relativa a deixar la carrera.

### GRADES\_SEM\_1 i GENDER



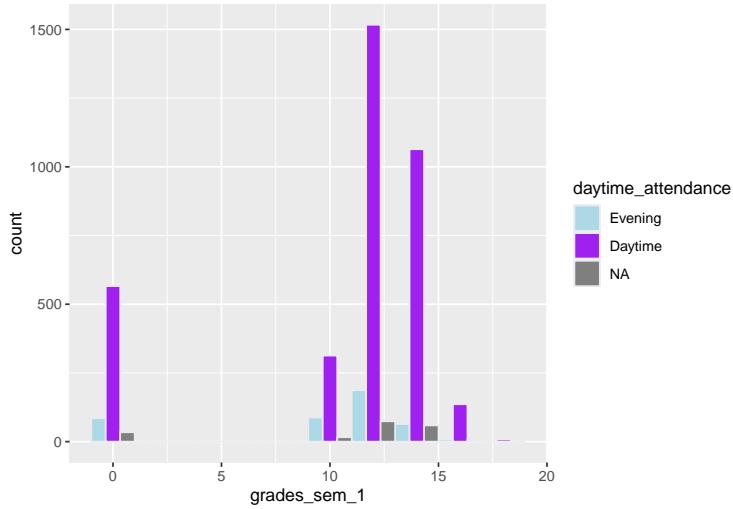
Les notes del primer semestre semblen distribuir-se igualment entre els dos gèneres. Sabent que el 0 significa no presentat s'observa que els homes tenen més tendència a no presentar-se que les dones, tenint en compte que hi ha més dones que homes.

### *GRADES\_SEM\_1 i SCHOLARSHIP*



Hi ha correlació entre treure notes més altes i tenir una “scholarship”. Si no es té beca és més probable no presentar-se.

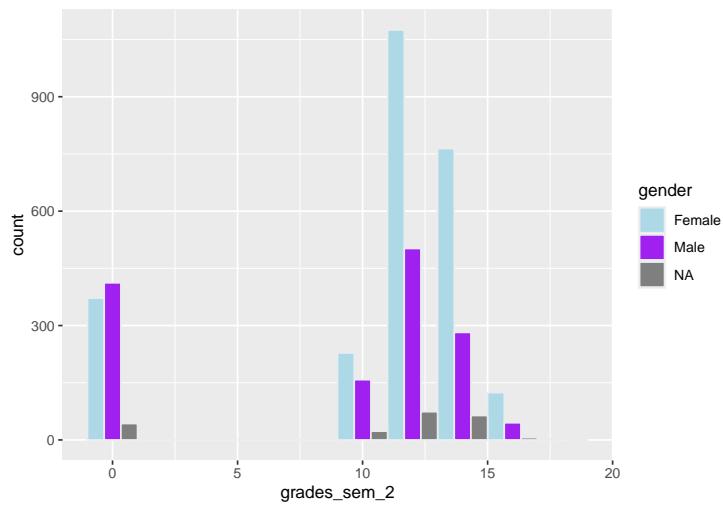
### *GRADES\_SEM\_1 i DAYTIME\_ATTENDANCE*



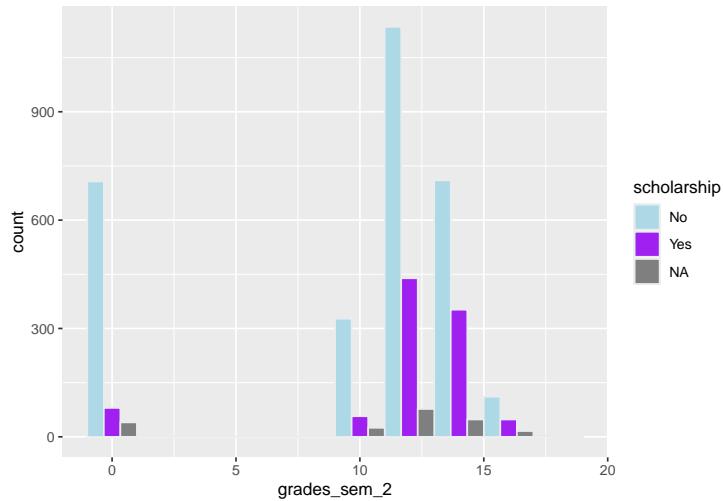
Les persones que assisteixen a classes de tarda tenen una mitjana de notes en el primer semestre més baixa que les persones que van a classes al matí.

Referent a les notes del segon semestre, es repliquen els mateixos patrons que per a les notes del primer.

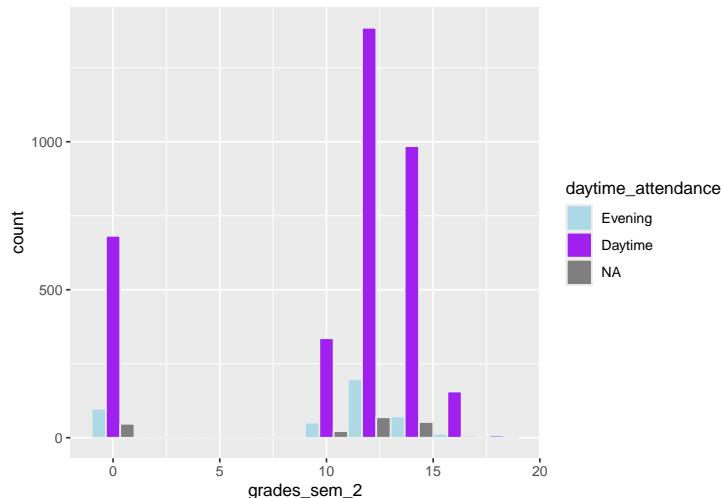
*GRADES\_SEM\_2* i GENDER



*GRADES\_SEM\_2* i SCHOLARSHIP



*GRADES\_SEM\_2* i DAYTIME\_ATTENDANCE



### 3.3 Preprocessing

En la primera part del preprocessing s'utilitzaran diferents mètodes per detectar outliers univariants i multivariants. En la segona part es tractaran els valors faltants de les variables categòriques i numèriques. L'objectiu és netejar la base de dades per poder garantir que les següents anàlisis siguin acurades.

#### 3.3.1 Detecció d'outliers

Es duu a terme una anàlisi per detectar outliers tant univariant com multivariant, ja que els valors atípics poden afectar potencialment a l'estimació dels paràmetres i a les conclusions extretes de l'informe.

##### Outliers multivariants

Per a fer la detecció d'outliers multivariants s'utilitzen les dades amb només variables numèriques i sense valors nuls, pel fet que es necessiten aquest tipus de dades per a dur a terme el mètode escollit: Distància de Mahalanobis. Aquesta mesura serveix per mesurar la distància entre dos vectors, on cada vector és una observació (individu). La distància de Mahalanobis es basa en la distància euclidiana i permet identificar combinacions atípiques de diverses variables.

El resultat de l'anàlisi d'outliers multivariant és que hi ha 97 files que en contenen algun. Com no es pot reconèixer quines són les variables d'aquestes files que provoquen el valor atípic, es decideix que les dades i files es queden tal com estan perquè convertir les 97 files és una gran pèrdua d'informació i podríem incorrèr en errors en les conclusions.

##### Outliers univariants de les variables numèriques

- VARIABLES ‘gdp’, ‘inflation\_rate’ I ‘unemployment\_rate’

Aquestes tres variables mostren dades macroeconòmiques dels anys en què van estudiar els alumnes. Com que la recollida de dades es va fer en deu anys, els valors s'han d'agrupar en 10 blocs corresponents als anys. No hi ha cap valor que quedi fora d'aquests blocs, per tant, es pot dir que no hi ha valors atípics.

- VARIABLE ‘age’

La variable de l'edat segueix una distribució exponencial. Es poden observar alguns valors atípics, però aquests no responen a errors en la recollida de dades, sinó a casos d'estudiants d'edat més avançada. No hi ha indicis que es tracti de cap error, ja que el valor màxim és de 70, una edat poc comuna però factible per a un estudiant. Així doncs, no és necessari eliminar aquests valors.

- VARIABLES ‘grades\_sem\_1’ I ‘grades\_sem\_2’

Aquestes dues variables indiquen les qualificacions dels alumnes. En el sistema universitari portuguès (on es desenvolupa l'estudi) les assignatures es qualifiquen del 0 al 20. Observant el rang de les dades es pot veure que no hi ha cap outlier.

- VARIABLES ‘no\_eval\_sem\_1’ I ‘no\_eval\_sem\_2’

Aquestes variables indiquen la quantitat de crèdits no evaluats de cada alumne. Mentre el valor sigui inferior al nombre de crèdits matriculats no hi ha indicis que hi hagi outliers.

### 3.3.2 Tractament de NA's

El tractament de NA's es duu a terme per variables categòriques i per variables numèriques. Per les variables categòriques es crea una categoria per agrupar els NA's, mentre que per les variables numèriques s'utilitzarà el mètode d'imputació MIMMI.

#### NA's en variables categòriques

Es crea una variable 'Unknown', que serà un nou nivell de les variables factors, que agrupa tots els NA's d'aquestes variables.

#### NA's en variables numèriques (MIMMI)

Per tractar els valors faltants de les variables numèriques es farà servir el mètode Mixed Intelligent Multivariate Missing Imputation (MIMMI). Aquest mètode permet imputar els NA's en bases de dades multivariantes.

Per aplicar el mètode de MIMMI a les dades s'ha treballat sobre el codi de Karina Gibert: *Gibert, K. (2014). Mixed intelligent-multivariate missing imputation. International Journal of Computer Mathematics, 91(1), 85-96 Barcelona May 2019*

Diem que k=3 perquè la població es divideix en 3 categories (Graduate, Dropout, Enrolled).

## 3.4 Anàlisi descriptiva després del preprocessing

### 3.4.1 Anàlisi univariant de les variables

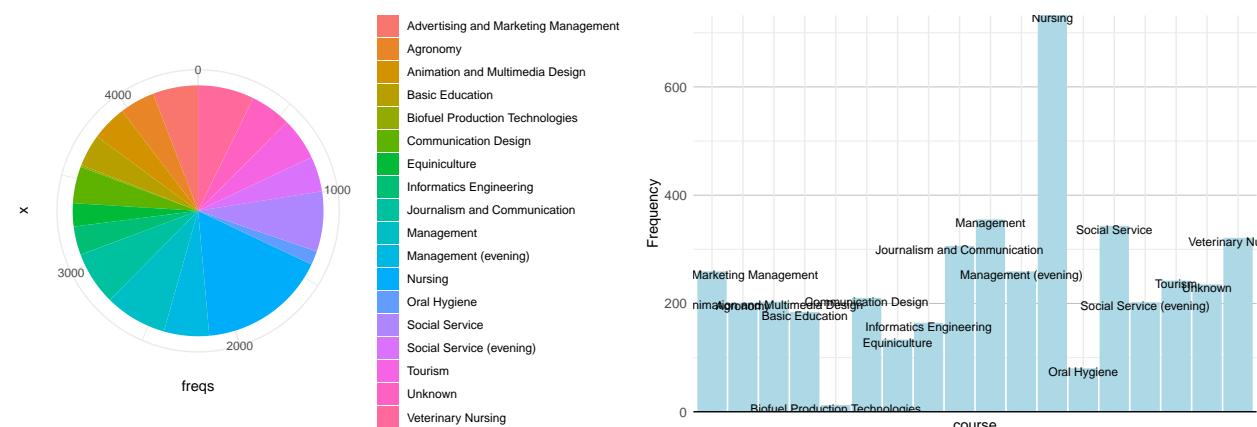
Per la part de l'anàlisi descriptiva després del preprocessing, en vista que els resultats són molt similars a aquells d'abans del preprocessing, no es farà una anàlisi en profunditat de cada gràfic, sinó que es remetrà a l'anàlisi d'abans del preprocessing per als gràfics que es presenten. A la part d'anàlisi descriptiva bivariant, en canvi, sí que s'estudiaran tots els gràfics.

#### Variables categòriques

S'ha afegit una variable nova, Unknown, que conté les dades faltants. Com que els missings s'han creat artificialment perquè representin només el 5% de la variable, l'addició de la variable Unknown no provoca canvis significatius en la descriptiva.

En les variables categòriques, la diferència més notòria és que les classes amb més percentatge es veuen reduïdes per la variable Unknown.

*COURSE*

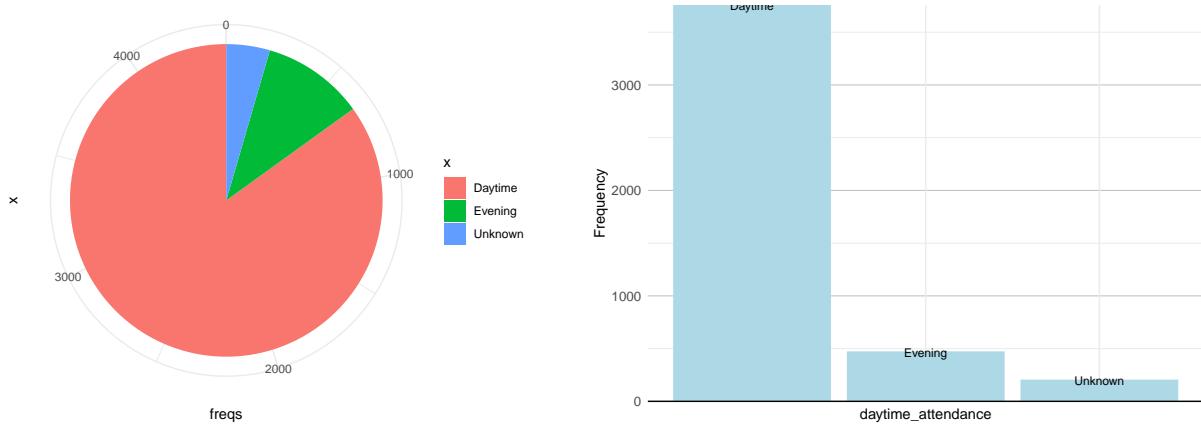


##

## -----

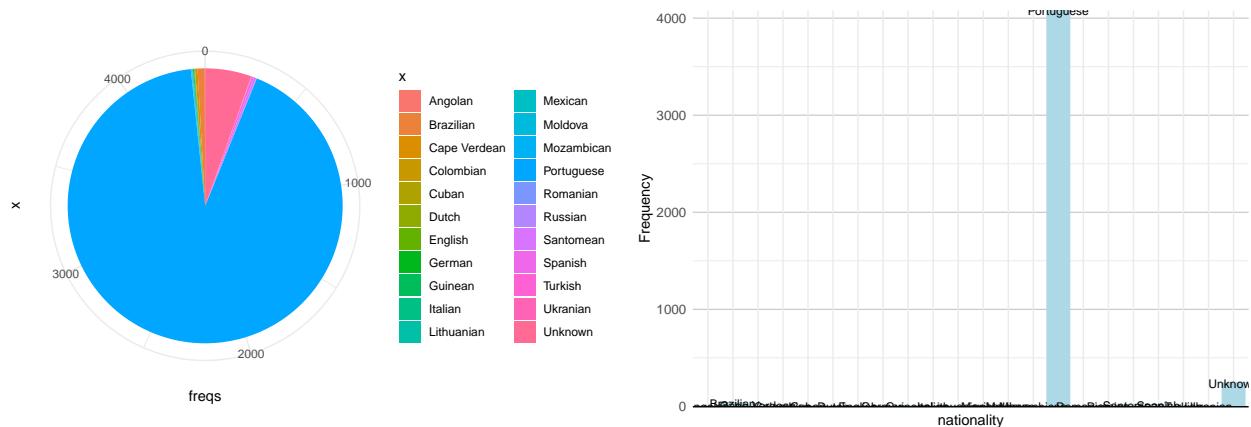
## Variable	Stats / Values	Freqs (% of Valid)	Missing
<hr/>			
## course\	1\. Advertising and Marketing\	258 ( 5.8%)\	0\
## [factor]	2\. Agronomy\	200 ( 4.5%)\	(0.0%)
##	3\. Animation and Multimedia \	202 ( 4.6%)\	
##	4\. Basic Education\	183 ( 4.1%)\	
##	5\. Biofuel Production Techno\	11 ( 0.2%)\	
##	6\. Communication Design\	209 ( 4.7%)\	
##	7\. Equiniculture\	132 ( 3.0%)\	
##	8\. Informatics Engineering\	163 ( 3.7%)\	
##	9\. Journalism and Communicat\	305 ( 6.9%)\	
##	10\. Management\	354 ( 8.0%)\	
##	[ 8 others ]	2407 (54.4%)	
<hr/>			

### DAYTIME\_ATTENDANCE



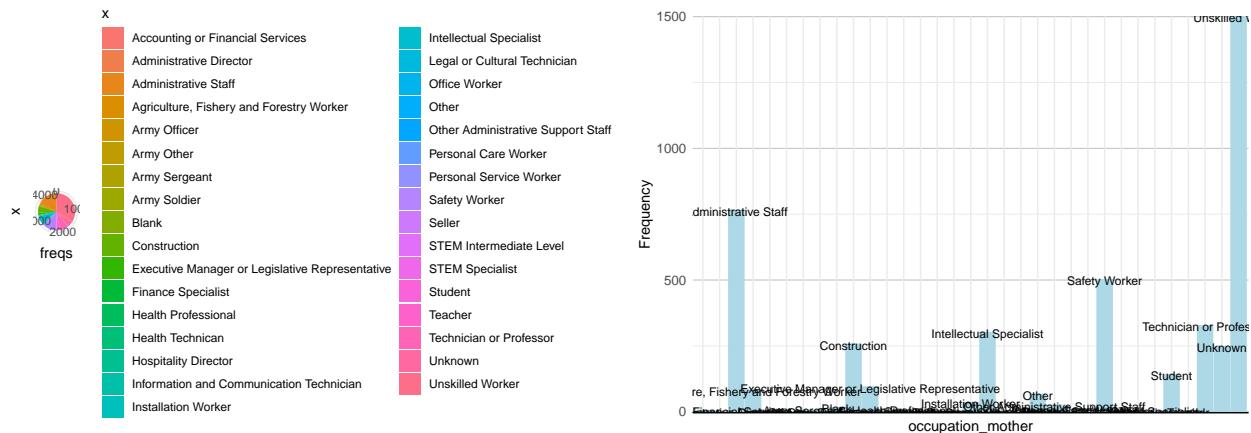
##	##	##	##
## Variable	Stats / Values	Freqs (% of Valid)	Missing
<hr/>			
## daytime_attendance\	1\. Daytime\	3758 (84.9%)\	0\
## [factor]	2\. Evening\	467 (10.6%)\	(0.0%)
##	3\. Unknown	199 ( 4.5%)	
<hr/>			

## NATIONALITY



```
##  
## -----  
## Variable      Stats / Values      Freqs (% of Valid)      Missing  
## -----  
## nationality\ 1\. Angolan\      2 ( 0.0%)\      0\  
## [factor]      2\. Brazilian\     37 ( 0.8%)\     (0.0%)  
##            3\. Cape Verdean\    13 ( 0.3%)\  
##            4\. Colombian\     1 ( 0.0%)\  
##            5\. Cuban\        1 ( 0.0%)\  
##            6\. Dutch\        1 ( 0.0%)\  
##            7\. English\      1 ( 0.0%)\  
##            8\. German\       2 ( 0.0%)\  
##            9\. Guinean\      5 ( 0.1%)\  
##           10\. Italian\     3 ( 0.1%)\  
##           [ 12 others ]    4358 (98.5%)  
## -----
```

## OCCUPATION\_MOTHER



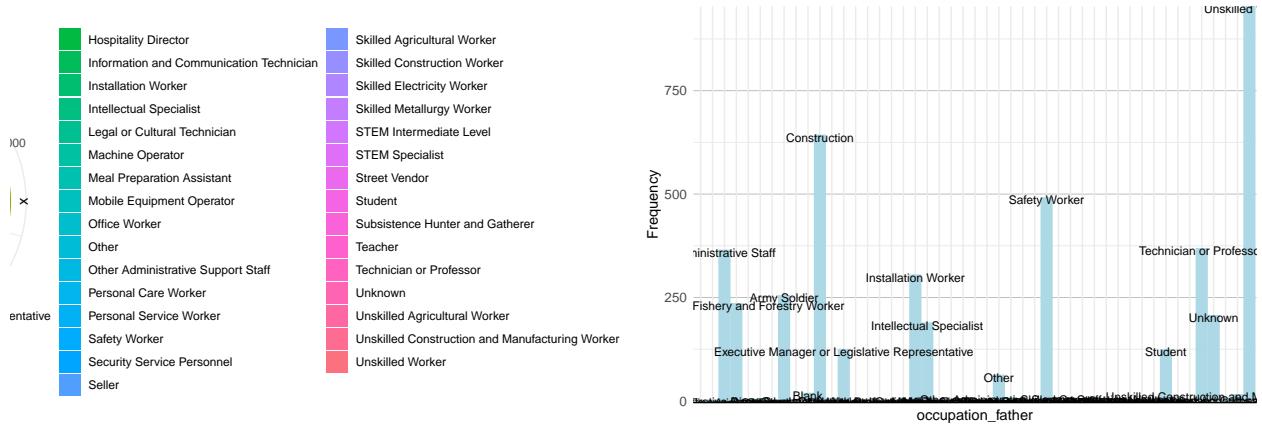
```
##  
## -----  
## Variable      Stats / Values      Freqs (% of Valid)      Missing  
## -----
```

```

## occupation_mother\ 1\. Accounting or Financial S\ 5 ( 0.1%)\ 0\
## [factor] 2\. Administrative Director\ 1 ( 0.0%)\ (0.0%)
## 3\. Administrative Staff\ 765 (17.3%)\ 
## 4\. Agriculture, Fishery and \ 82 ( 1.9%)\ 
## 5\. Army Officer\ 2 ( 0.0%)\ 
## 6\. Army Other\ 1 ( 0.0%)\ 
## 7\. Army Sergeant\ 7 ( 0.2%)\ 
## 8\. Army Soldier\ 3 ( 0.1%)\ 
## 9\. Blank\ 15 ( 0.3%)\ 
## 10\. Construction\ 256 ( 5.8%)\ 
## [ 23 others ] 3287 (74.3%)
## -----

```

*OCCUPATION\_FATHER*

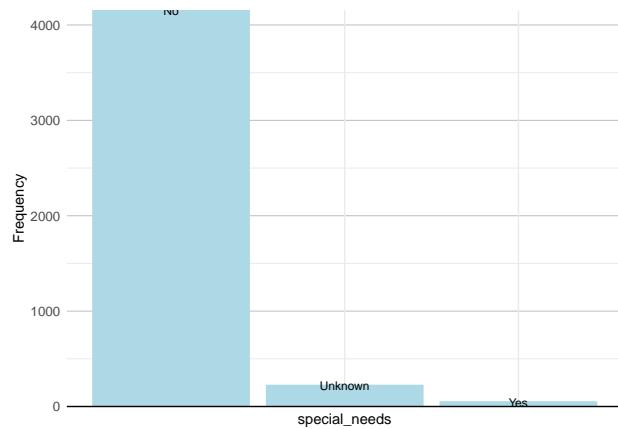
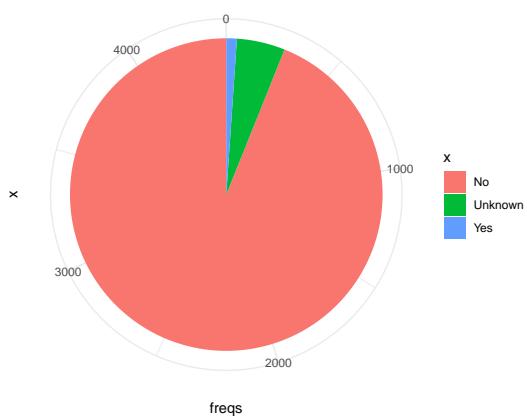


```

## 
## -----
## Variable           Stats / Values          Freqs (% of Valid)  Missing
## -----
## occupation_father\ 1\. Accounting or Financial S\ 1 ( 0.0%)\ 0\
## [factor]           2\. Administrative Director\ 2 ( 0.0%)\ (0.0%)
##                   3\. Administrative Staff\ 364 ( 8.2%)\ 
##                   4\. Agriculture, Fishery and \ 235 ( 5.3%)\ 
##                   5\. Army Officer\ 1 ( 0.0%)\ 
##                   6\. Army Other\ 4 ( 0.1%)\ 
##                   7\. Army Sergeant\ 2 ( 0.0%)\ 
##                   8\. Army Soldier\ 254 ( 5.7%)\ 
##                   9\. Assembly Worker\ 1 ( 0.0%)\ 
##                   10\. Blank\ 18 ( 0.4%)\ 
##                   [ 37 others ] 3542 (80.1%)
## 

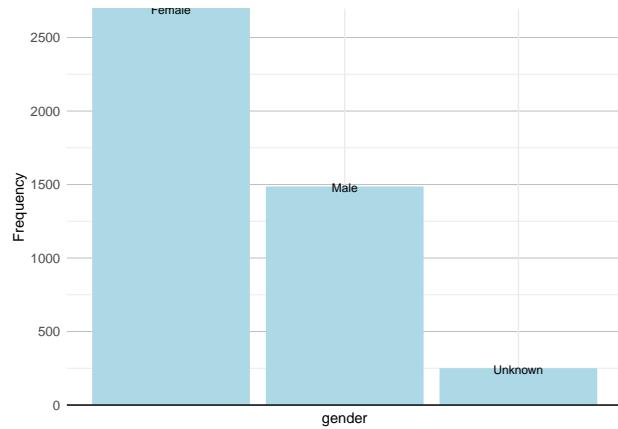
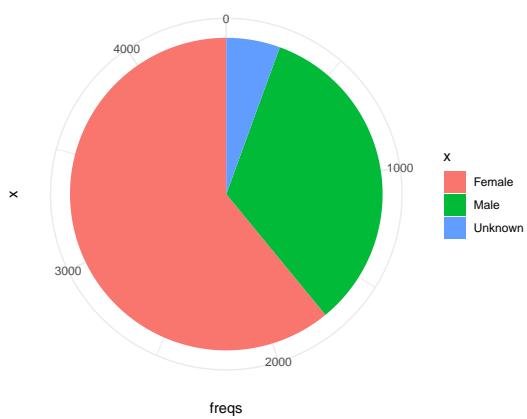
```

## SPECIAL\_NEEDS



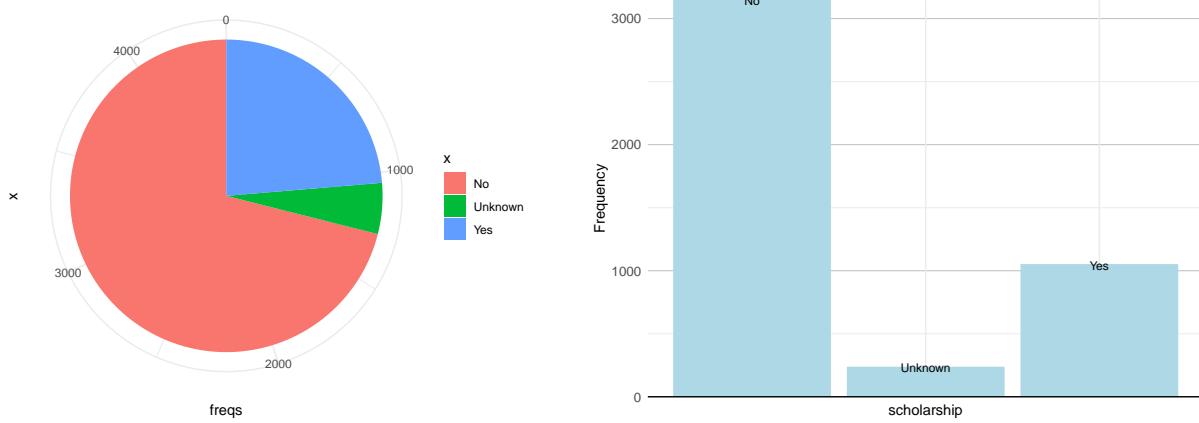
```
##  
## -----  
## Variable      Stats / Values   Freqs (% of Valid)  Missing  
## -----  
## special_needs\\ 1\\. No\\       4155 (93.9%)\\    0\\  
## [factor]        2\\. Unknown\\  221 ( 5.0%)\\     (0.0%)  
##                  3\\. Yes        48 ( 1.1%)  
## -----
```

## GENDER



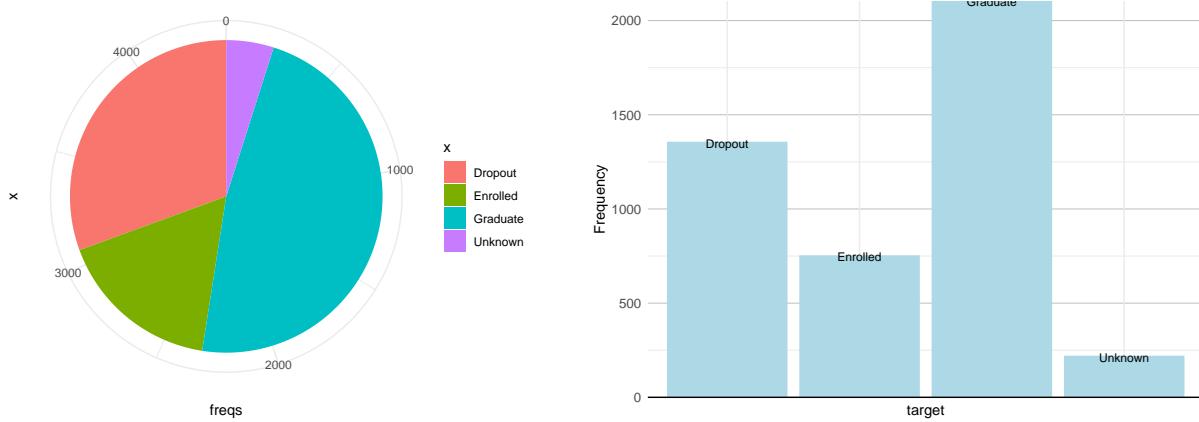
```
##  
## -----  
## Variable      Stats / Values   Freqs (% of Valid)  Missing  
## -----  
## gender\\       1\\. Female\\    2696 (60.9%)\\    0\\  
## [factor]        2\\. Male\\      1482 (33.5%)\\    (0.0%)  
##                  3\\. Unknown    246 ( 5.6%)  
## -----
```

## SCHOLARSHIP



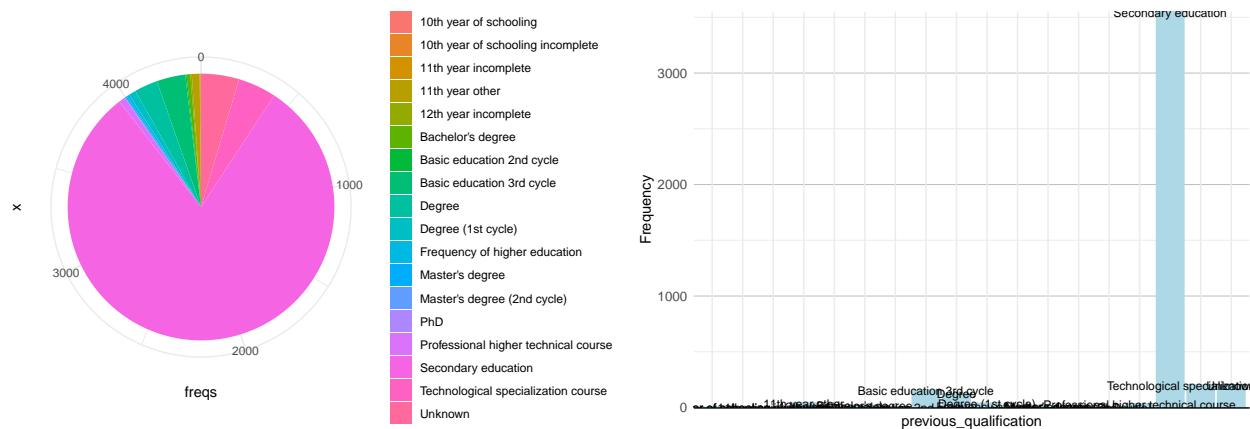
```
##  
## -----  
## Variable      Stats / Values   Freqs (% of Valid)  Missing  
## -----  
## scholarship\  1\. No\        3144 (71.1%)\  
## [factor]       2\. Unknown\    233 ( 5.3%)\  
##                 3\. Yes        1047 (23.7%)  
## -----
```

## TARGET



```
##  
## -----  
## Variable      Stats / Values   Freqs (% of Valid)  Missing  
## -----  
## target\       1\. Dropout\    1353 (30.6%)\  
## [factor]       2\. Enrolled\    750 (17.0%)\  
##                 3\. Graduate\   2104 (47.6%)\  
##                 4\. Unknown     217 ( 4.9%)  
## -----
```

## PREVIOUS\_QUALIFICATION

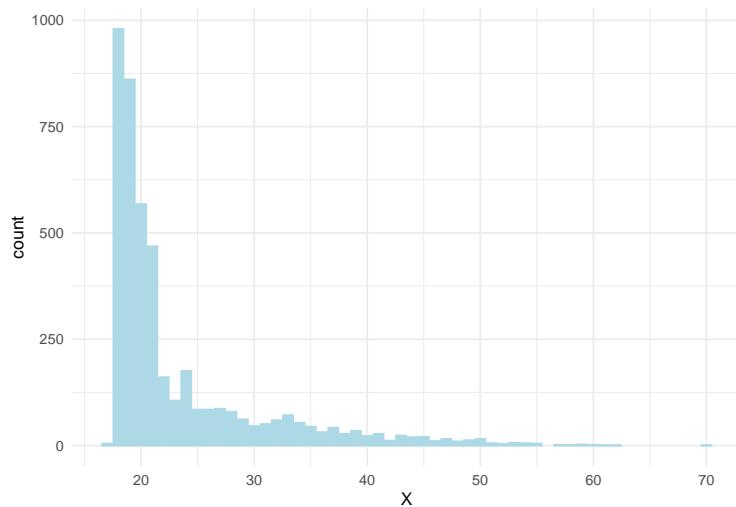


```
##  
## -----  
## Variable           Stats / Values      Freqs (% of Valid)  Missing  
## -----  
## previous_qualification\  
## [factor]          1\. 10th year of schooling\ 1 ( 0.0%)\  
##                      2\. 10th year of schooling in\ 2 ( 0.0%)\  
##                      3\. 11th year incomplete\ 4 ( 0.1%)\  
##                      4\. 11th year other\ 42 ( 0.9%)\  
##                      5\. 12th year incomplete\ 8 ( 0.2%)\  
##                      6\. Bachelor's degree\ 20 ( 0.5%)\  
##                      7\. Basic education 2nd cycle\ 7 ( 0.2%)\  
##                      8\. Basic education 3rd cycle\ 154 ( 3.5%)\  
##                      9\. Degree\ 125 ( 2.8%)\  
##                     10\. Degree (1st cycle)\ 38 ( 0.9%)\  
##                     [ 8 others ]        4023 (90.9%)  
## -----
```

## Variables numèriques

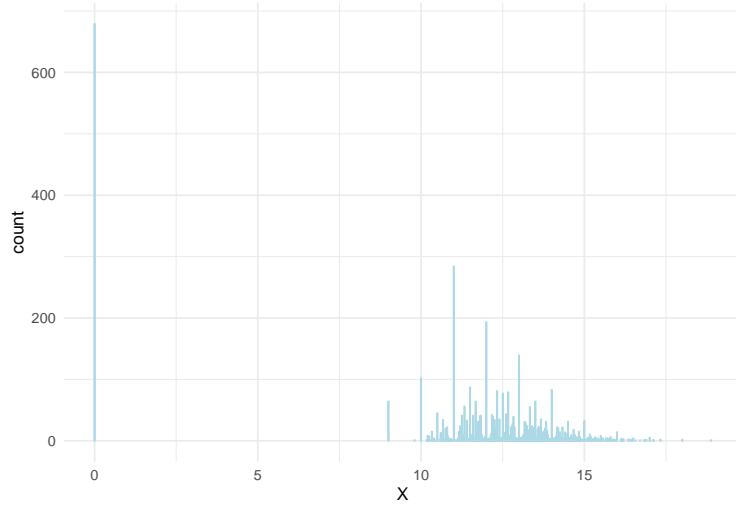
En les variables numèriques tampoc s'observen canvis significatius. El més destacable és que la mediana de la taxa d'inflació ha baixat de 1.4% a 1% i la mitjana del PIB ha variat de 0.003% a 0.011%.

*AGE*



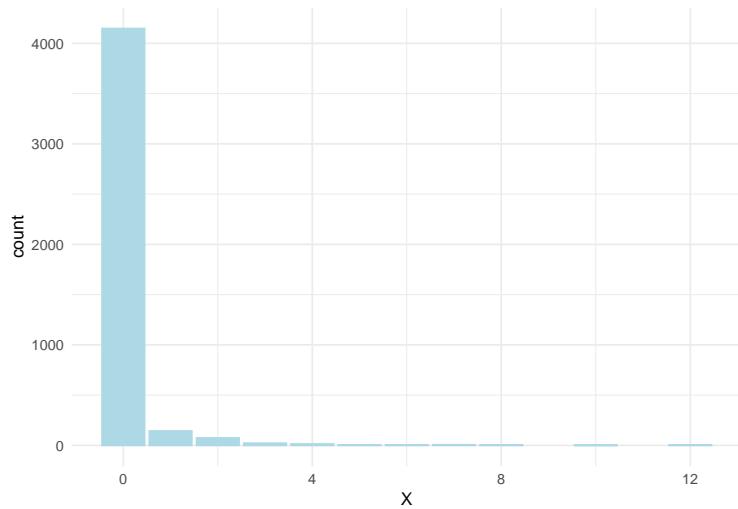
```
##  
## | N.Valid | Min | Q1 | Median | Mean | Std.Dev | Q3 | Max | IQR |  
## |:-----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|  
## | 4424.000 | 17.000 | 19.000 | 20.000 | 23.244 | 7.426 | 24.000 | 70.000 | 5.000 |
```

*GRADES\_SEM\_1*



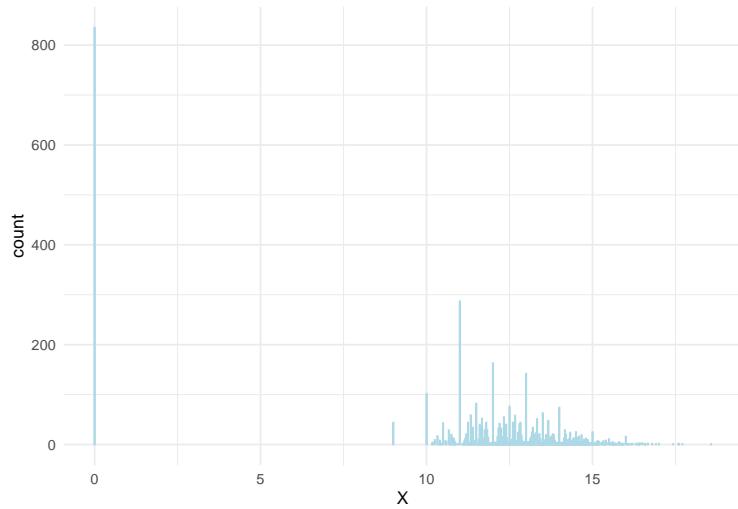
```
##  
## | N.Valid | Min | Q1 | Median | Mean | Std.Dev | Q3 | Max | IQR |  
## |:-----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|  
## | 4424.000 | 0.000 | 11.000 | 12.170 | 10.629 | 4.717 | 13.330 | 18.880 | 2.330 |
```

### *NO\_EVAL\_SEM\_1*



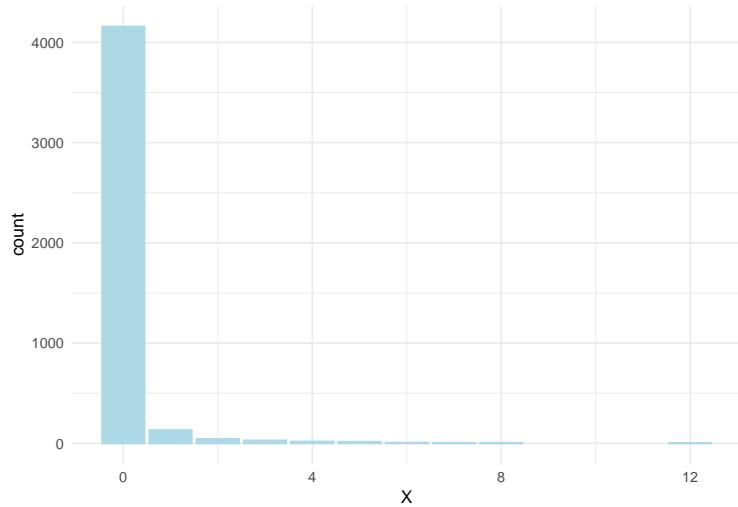
```
##  
## | N.Valid | Min | Q1 | Median | Mean | Std.Dev | Q3 | Max | IQR |  
## |:-----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|  
## | 4424.000 | 0.000 | 0.000 | 0.000 | 0.130 | 0.677 | 0.000 | 12.000 | 0.000 |
```

### *GRADES\_SEM\_2*



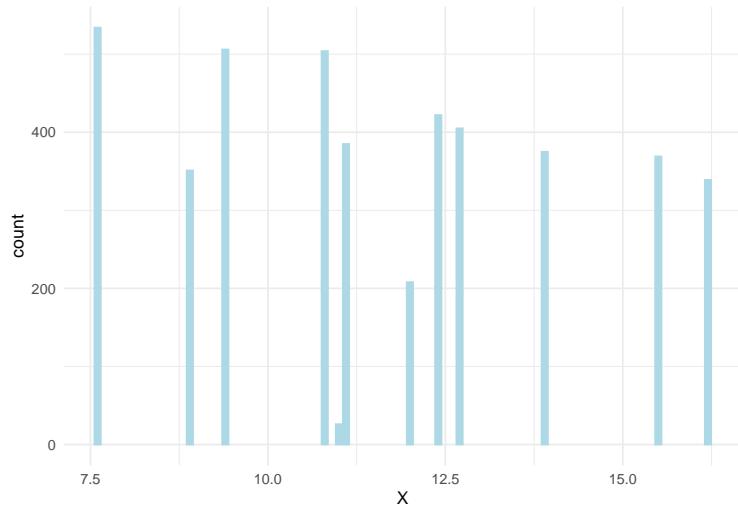
```
##  
## | N.Valid | Min | Q1 | Median | Mean | Std.Dev | Q3 | Max | IQR |  
## |:-----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|  
## | 4424.000 | 0.000 | 10.670 | 12.070 | 10.222 | 5.101 | 13.300 | 18.570 | 2.630 |
```

### *NO\_EVAL\_SEM\_2*



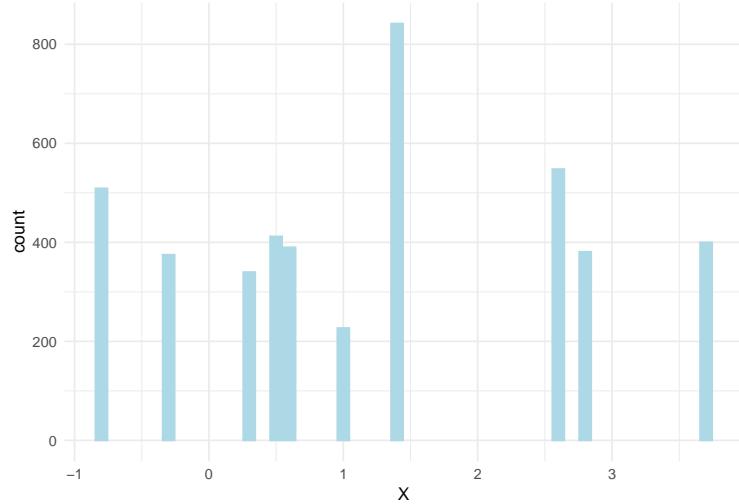
```
##  
## | N.Valid | Min | Q1 | Median | Mean | Std.Dev | Q3 | Max | IQR |  
## |:-----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|  
## | 4424.000 | 0.000 | 0.000 | 0.000 | 0.141 | 0.731 | 0.000 | 12.000 | 0.000 |
```

### *UNEMPLOYMENT\_RATE*



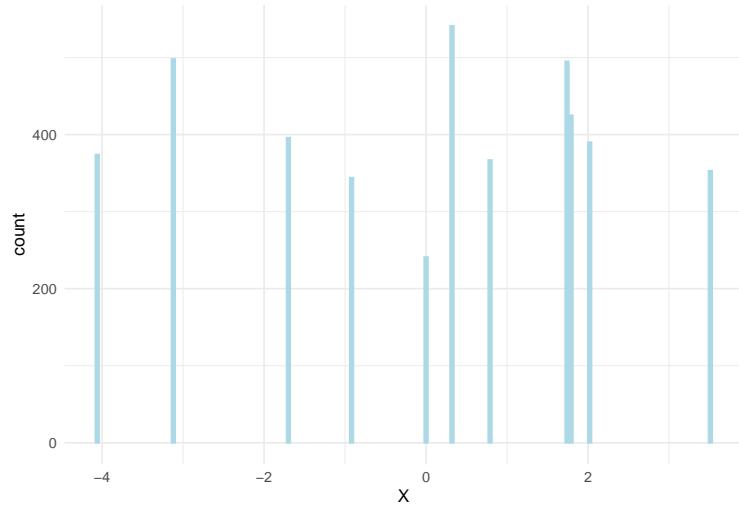
```
##  
## | N.Valid | Min | Q1 | Median | Mean | Std.Dev | Q3 | Max | IQR |  
## |:-----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|  
## | 4424.000 | 7.600 | 9.400 | 11.100 | 11.582 | 2.584 | 12.700 | 16.200 | 3.300 |
```

### *INFLATION\_RATE*



```
##  
## | N.Valid | Min | Q1 | Median | Mean | Std.Dev | Q3 | Max | IQR |  
## |:-----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|  
## | 4424.000 | -0.800 | 0.300 | 1.000 | 1.221 | 1.352 | 2.600 | 3.700 | 2.300 |
```

### *GDP*

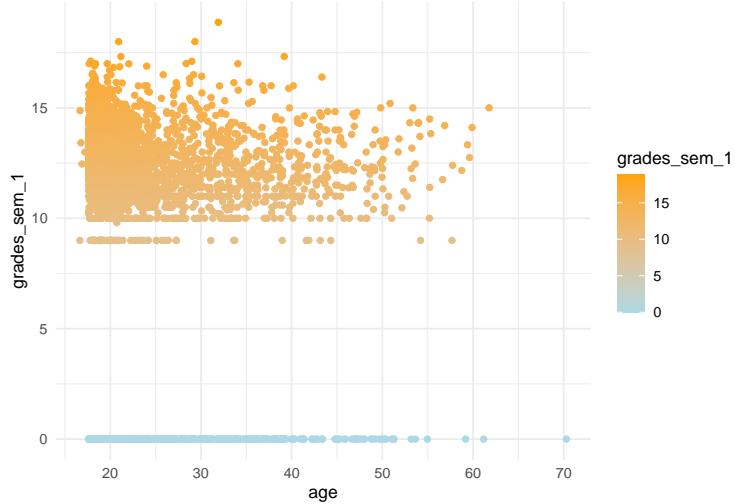


```
##  
## | N.Valid | Min | Q1 | Median | Mean | Std.Dev | Q3 | Max | IQR |  
## |:-----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|:----:|  
## | 4424.000 | -4.060 | -1.700 | 0.320 | 0.011 | 2.206 | 1.790 | 3.510 | 3.490 |
```

### 3.4.2 Anàlisi bivariant

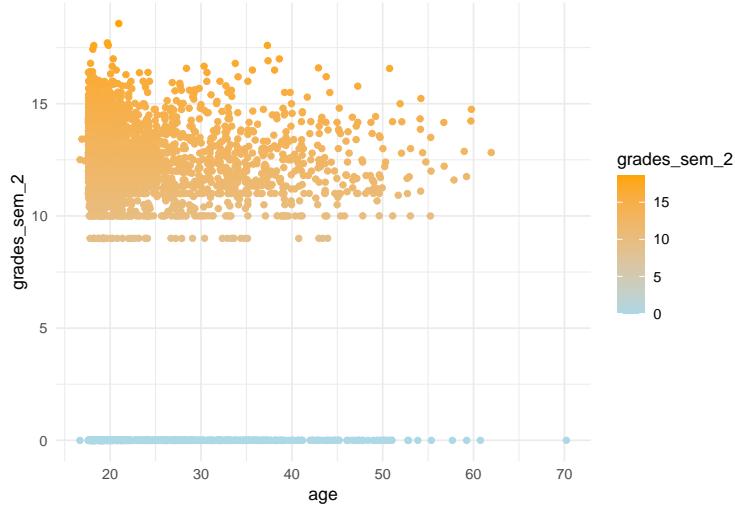
#### Dues variables numèriques

*AGE* i *GRADES\_SEM\_1*



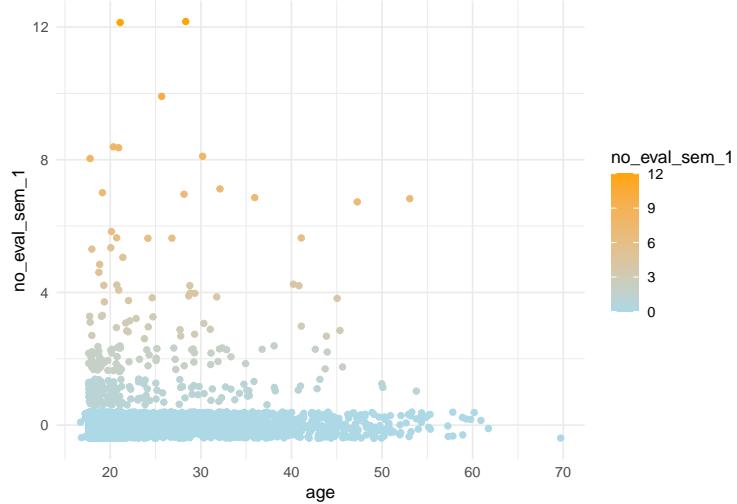
En funció de l'edat, s'observa que no hi ha diferències significatives a la nota del primer semestre. És possible que, en entrar en una edat més avançada (dels 50 endavant) hi hagi menys tendència a les notes excel·lents.

*AGE* i *GRADES\_SEM\_2*



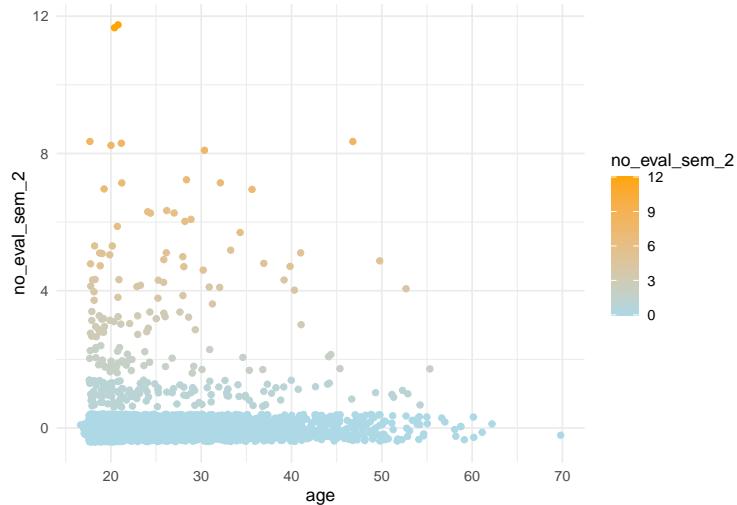
Per a les notes del segon semestre, s'observa el mateix que amb les notes del primer.

### AGE i NO\_EVAL\_SEM\_1



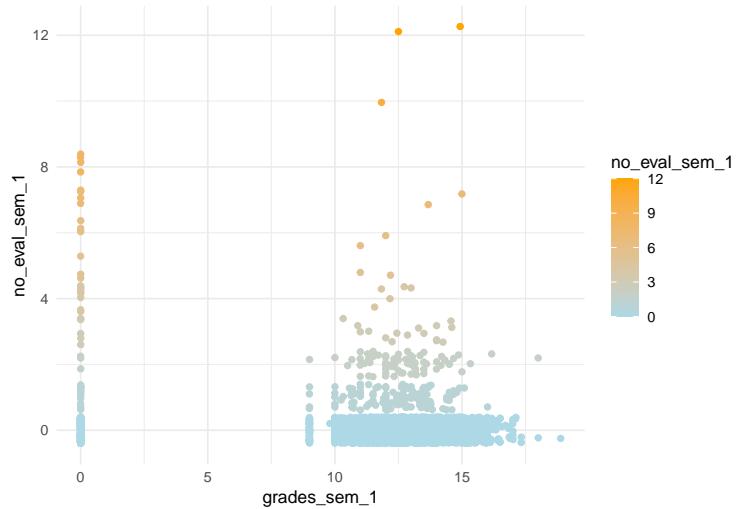
L'edat no sembla significativa respecte al nombre d'unitats no avaluades al primer semestre.

### AGE i NO\_EVAL\_SEM\_2



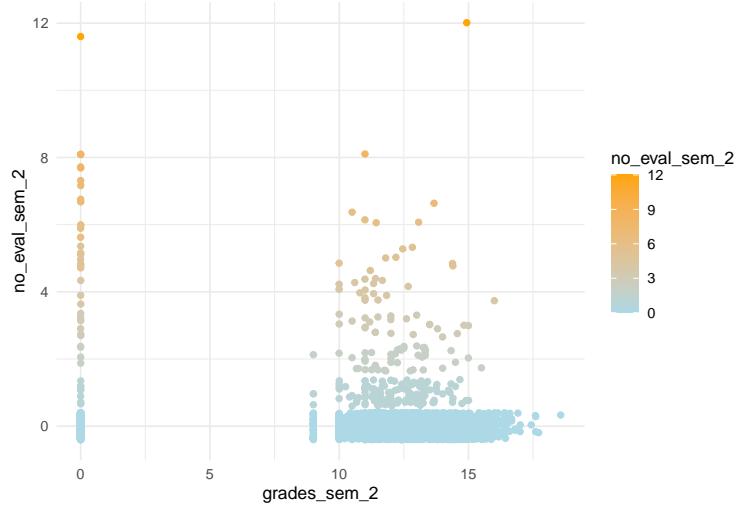
Igual que pel primer semestre, l'edat no sembla significativa respecte al nombre d'unitats no avaluades al segon semestre.

*GRADES\_SEM\_1* i *NO\_EVAL\_SEM\_1*



No sembla haver-hi correlació entre les notes del primer semestre i el nombre d'unitats no avaluades al primer semestre.

*GRADES\_SEM\_2* i *NO\_EVAL\_SEM\_2*

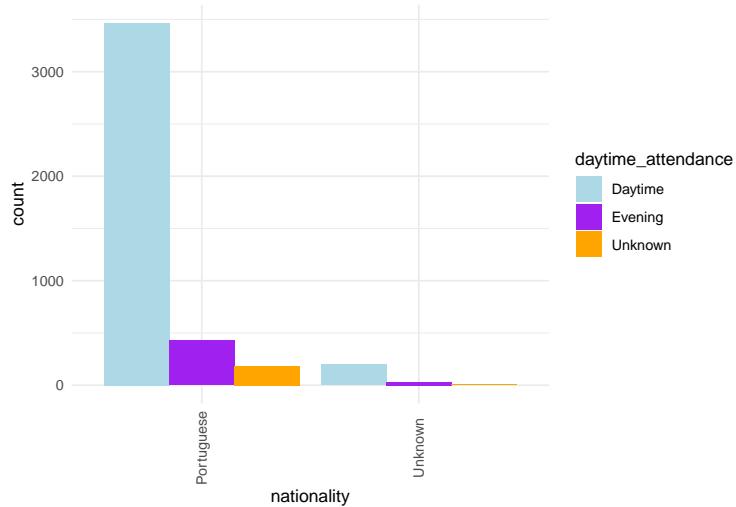


No sembla haver-hi correlació entre les notes del segon semestre i el nombre d'unitats no avaluades al segon semestre.

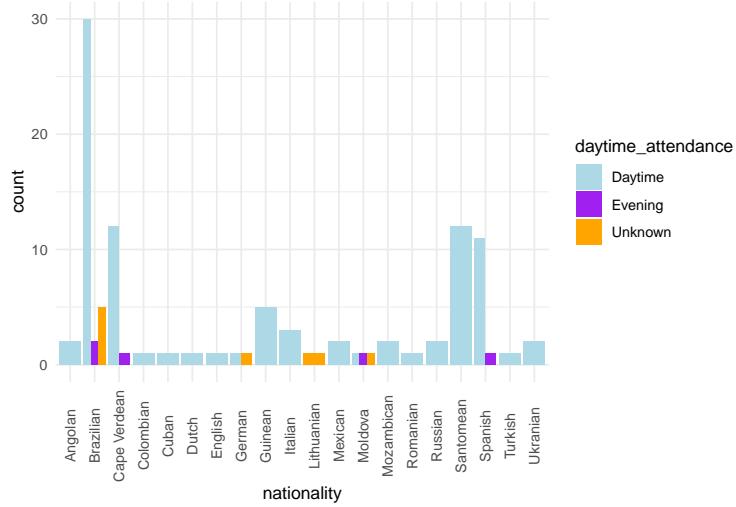
**Dues variables categòriques**

*NATIONALITY* i *DAYTIME\_ATTENDANCE*

Per fer l'anàlisi de la variable “nationality”, se separaren els casos de persones portugueses o de nacionalitat desconeguda respecte dels altres casos, ja que la diferència en nombre de casos per les dues classificacions no permet veure bé les dades de les persones no portugueses.

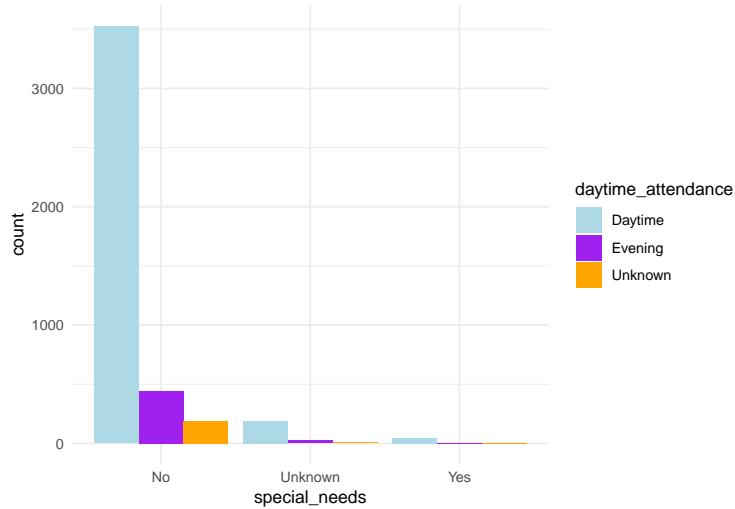


Pel cas de persones portugueses o de nacionalitat desconeguda, s'observa una distribució similar en assistir al matí, a la tarda o no haver-hi dades al respecte.



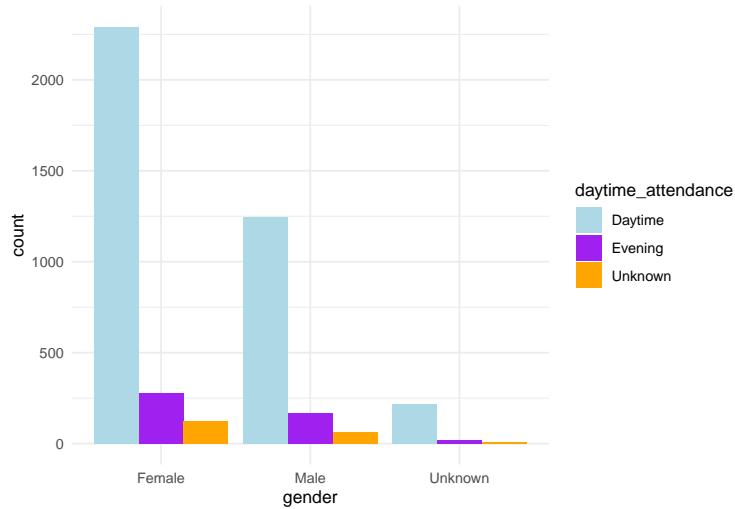
Pel cas de les persones no portugueses, el nombre reduït de casos no permet veure bé la distribució, però la majoria de les persones va al matí, igual que al cas de les persones portugueses o de nacionalitat desconeguda.

*SPECIAL\_NEEDS i DAYTIME\_ATTENDANCE*



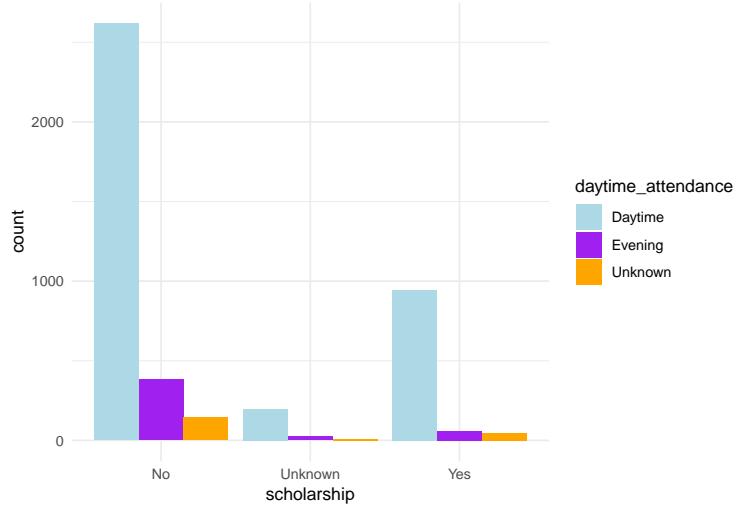
Sembla que independentment de si es tenen necessitats especials o no, o si no se sap, la distribució d'anar al dia, a la tarda o que no hi hagi dades és la mateixa que abans.

*GENDER i DAYTIME\_ATTENDANCE*



Independentment del gènere, la majoria de les persones van al matí.

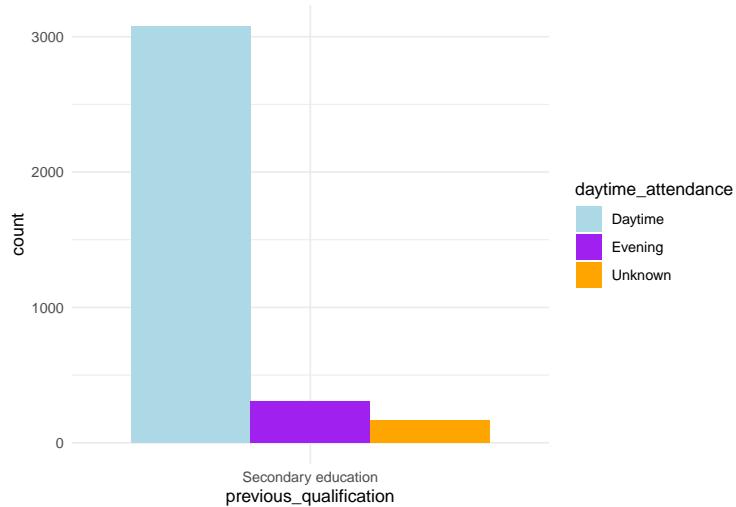
### *SCHOLARSHIP i DAYTIME\_ATTENDANCE*



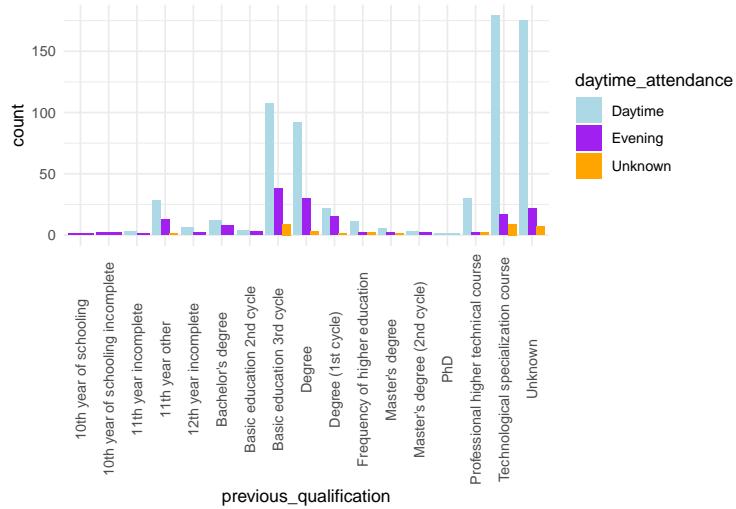
Sembla que les persones amb beca van menys a classe a la tarda que les persones sense beca, relativament a la quantitat de persones que hi ha de cada tipus.

### *PREVIOUS\_QUALIFICATION i DAYTIME\_ATTENDANCE*

Per l'anàlisi de dades que inclogui la variable de la qualificació prèvia, se separaran els casos de persones amb educació secundària respecte de les altres, per poder fer una visualització millor de les dades.

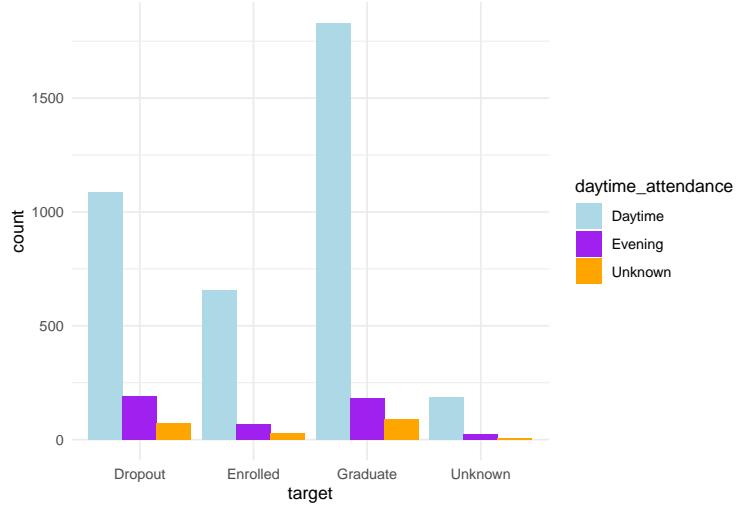


La distribució de matí, tarda i sense dades és la mateixa que s'ha anat observant pel cas de l'educació secundària.



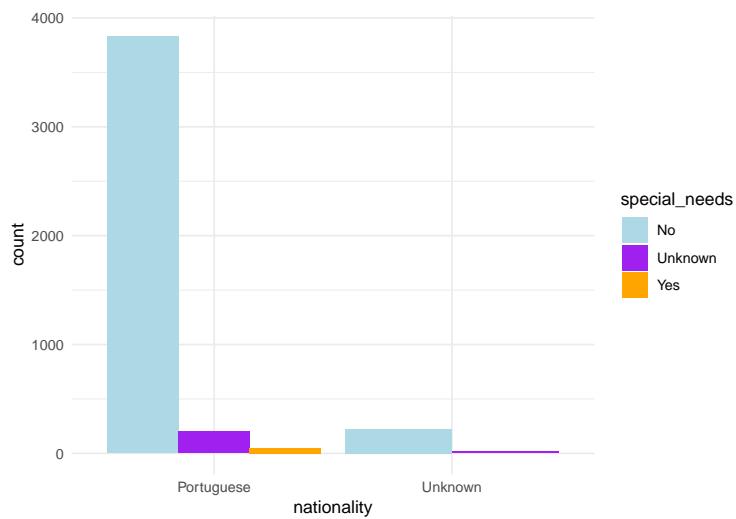
Per als altres casos, es pot observar que hi ha grups que tenen més tendència de l'usual a anar a la tarda, com ara les persones amb un grau o aquelles amb educació bàsica de tercer cicle. Això pot ser perquè aquests són grups que tenen més probabilitat de treballar al matí.

#### TARGET i DAYTIME\_ATTENDANCE

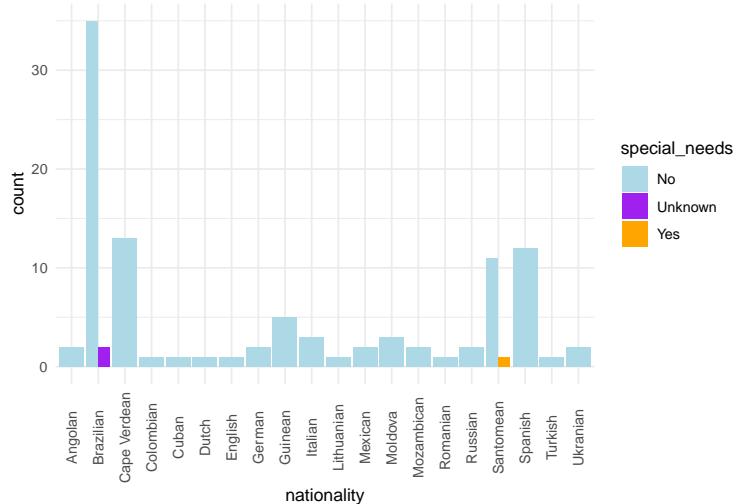


Les persones que van deixar el grau tenien més tendència a anar a classe a la tarda que les persones que es van graduar. Es pot observar que aproximadament és el doble de tendència.

## NATIONALITY i SPECIAL\_NEEDS

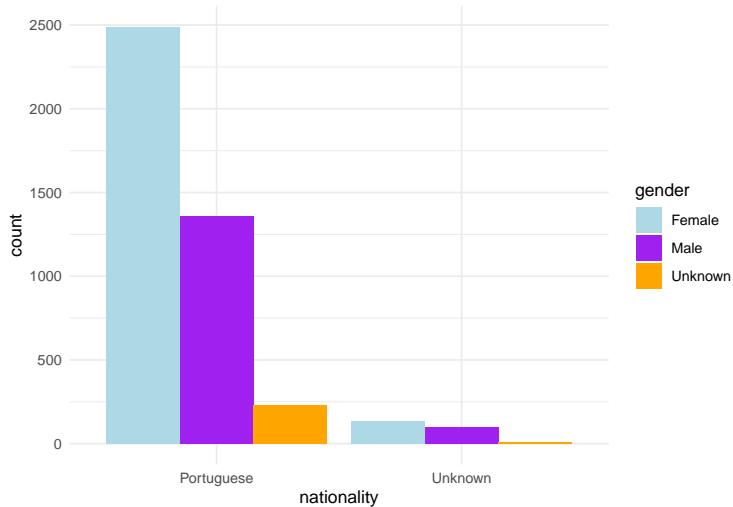


Les persones portugueses majoritàriament no tenen necessitats especials, tot i que es pot veure que aproximadament unes 200 personnes portugueses de la mostra sí que en tenen.

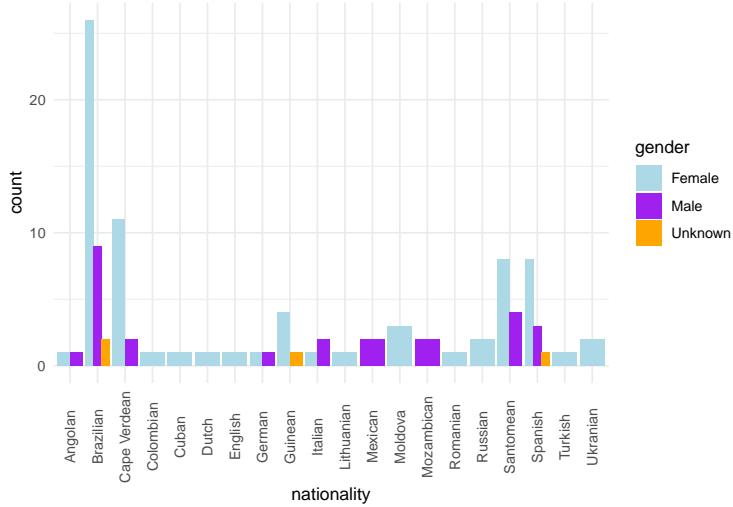


Per a les altres nacionalitats, no es pot saber significativament la probabilitat de tenir necessitats especials, ja que la mida de la mostra és molt petita. Es pot observar, però, que la majoria d'estudiants de nacionalitat no portuguesa no tenen necessitats especials a la mostra.

## NATIONALITY i GENDER

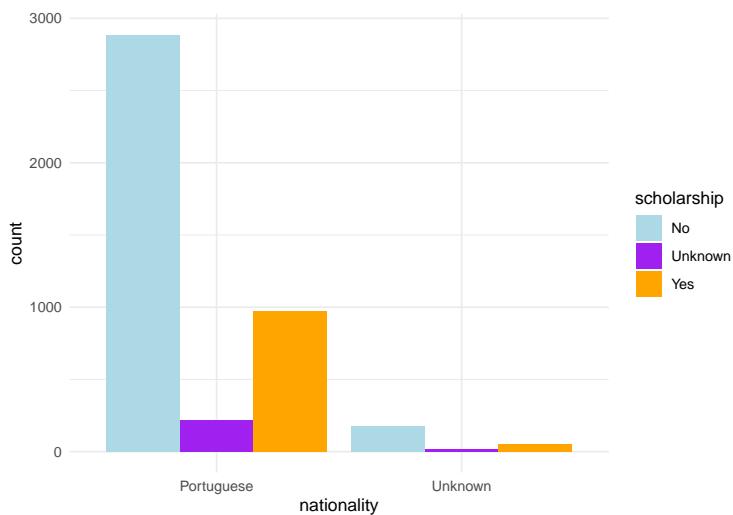


Es pot observar que pels casos de nacionalitat portuguesa, aproximadament hi ha el doble de dones que d'homes estudiant. D'una proporció significativa de persones no se sap el gènere. Pels casos de nacionalitat desconeguda, els gèneres estan més igualats, tot i que no del tot.

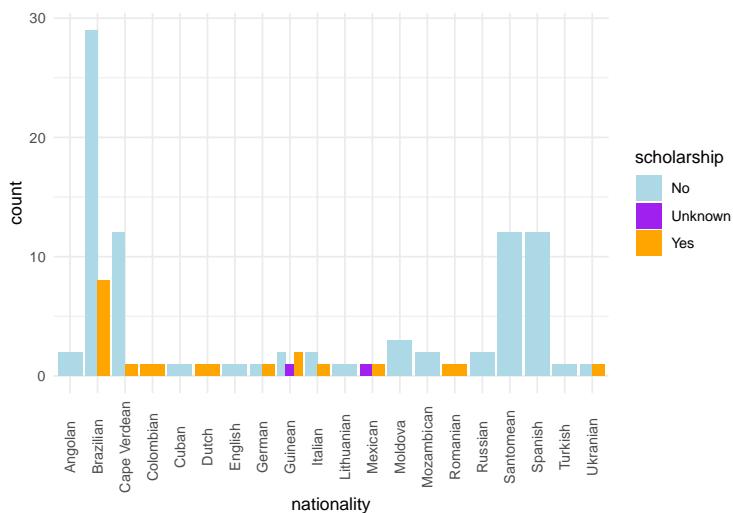


A les altres nacionalitats també s'observa més tendència a ser estudiant dona que home.

## NATIONALITY i SCHOLARSHIP

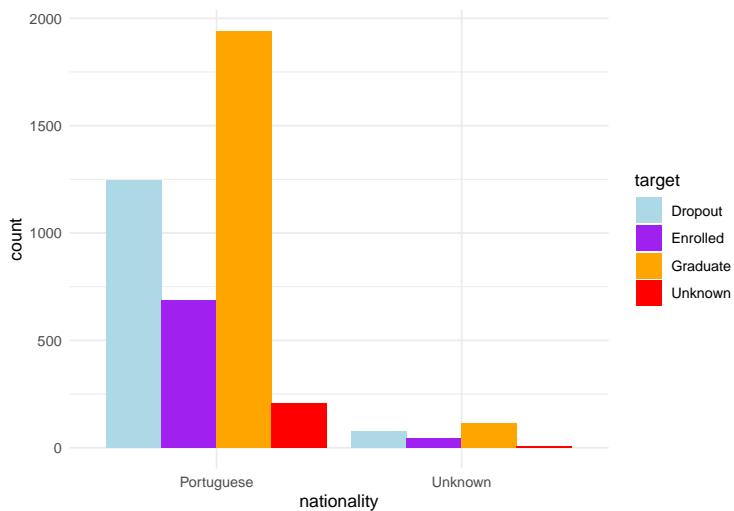


Per a les personnes portugueses, la probabilitat de tenir beca és aproximadament un quart. Pels casos de nacionalitat desconeguda s'observa una tendència similar.

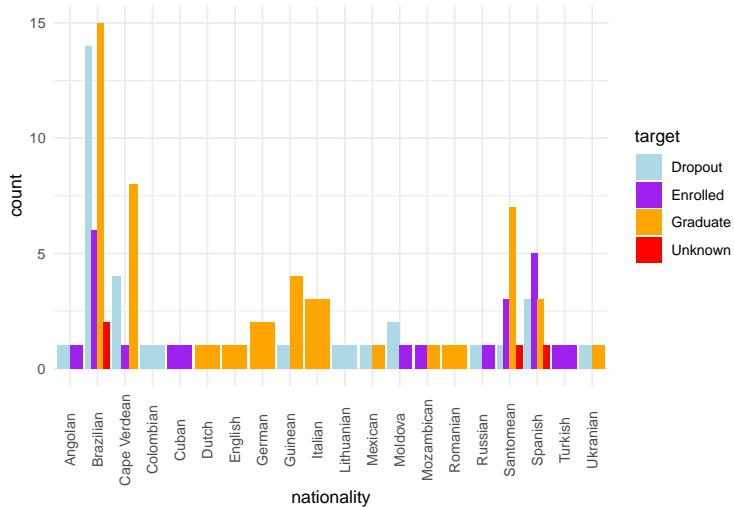


Per a les personnes no portugueses, no hi ha prou dades per a extreure conclusions significatives, però s'observa una tendència similar.

## NATIONALITY i TARGET

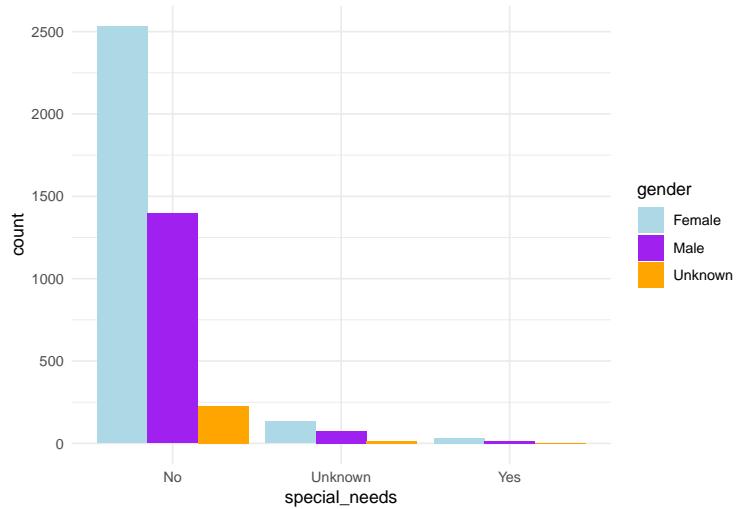


La probabilitat de graduar-se en persones portugueses és aproximadament un 140% de la probabilitat de deixar la carrera.



Amb el nombre baix de dades de nacionalitats no portugueses, no es pot saber bé la distribució, però sembla que es replica l'anterior.

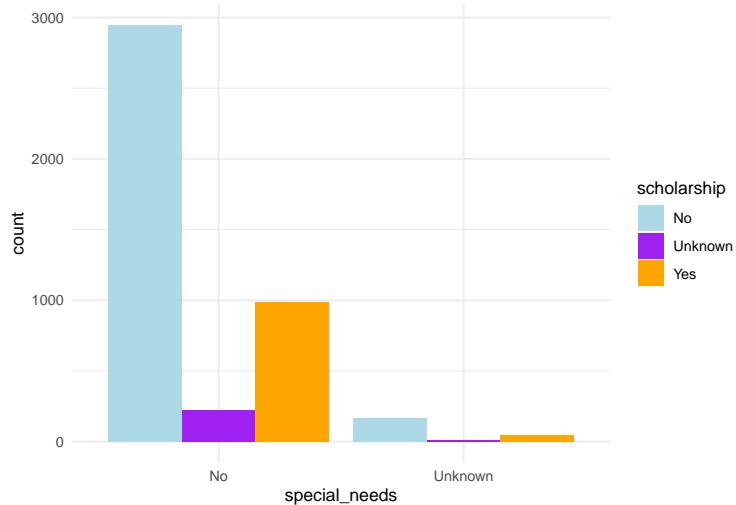
### *SPECIAL\_NEEDS i GENDER*



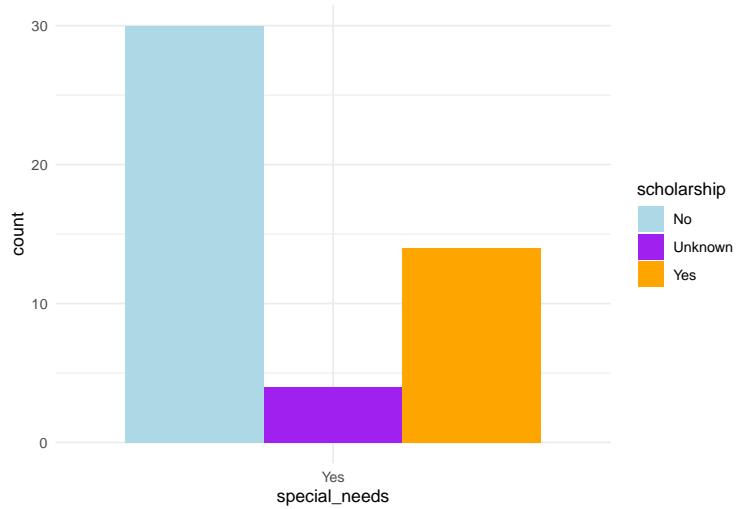
El gènere no sembla rellevant respecte de tenir necessitats especials o no.

### *SPECIAL\_NEEDS i SCHOLARSHIP*

Per als gràfics, observant el fet de tenir necessitats especials o no junt amb tenir scholarship o no, se separaren la categoria de tenir necessitats especials respecte de no tenir-ne o que no se sàpiga per poder visualitzar millor els resultats.



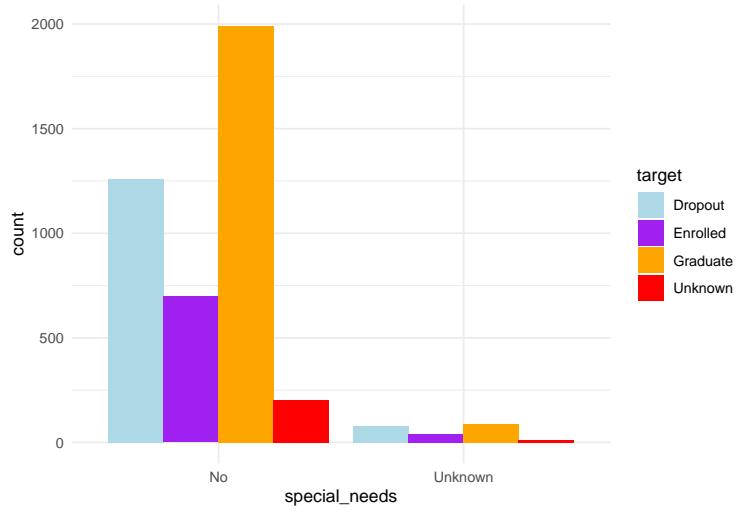
Per les persones sense necessitats especials, la probabilitat de tenir scholarship és aproximadament un quart. Sembla replicar-se per als casos desconeguts.



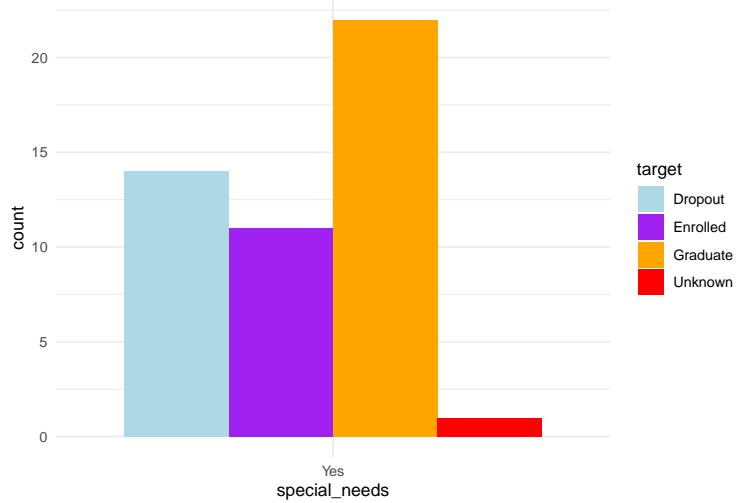
Per a les persones amb necessitats especials, la probabilitat de tenir beca sembla més aviat un terç. Tot i això, la mida mostra reduïda implica que el resultat no és significatiu.

Per a l'anàlisi conjunta de tenir necessitats especials o no i graduar-se o no, se separaren els casos de necessitats especials i els que no, per poder observar millor els resultats.

#### *SPECIAL\_NEEDS i TARGET*



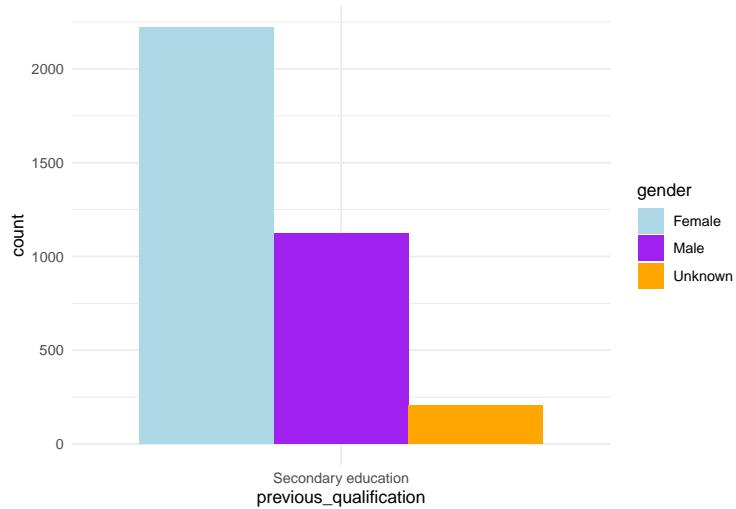
Sense necessitats especials, la probabilitat de graduar-se és aproximadament un 140% de la de deixar la carrera, altre cop. Pel cas de persones que no se sap si tenen necessitats especials o no, sembla que la probabilitat és la mateixa de graduar-se que de deixar-ho. Tot i això, la mida mostra és molt reduïda.



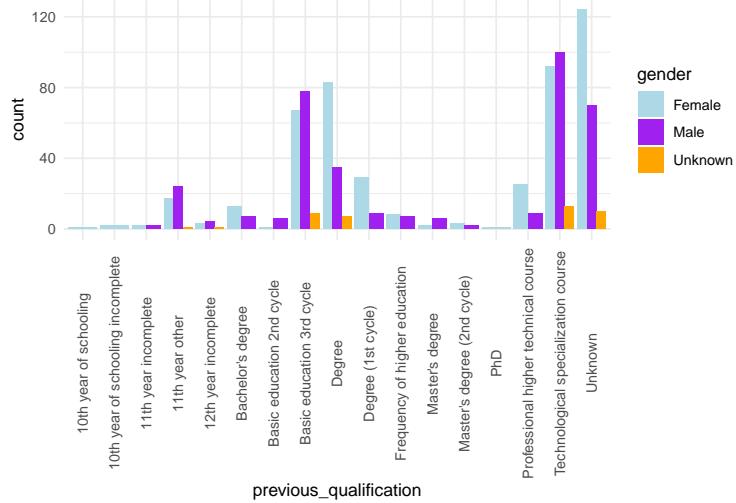
Per a les persones amb necessitats especials, s'observa la mateixa tendència que sense necessitats especials.

#### *PREVIOUS\_QUALIFICATION i GENDER*

Per l'anàlisi de gènere segons qualificació prèvia, se separaran els casos d'educació secundària respecte dels altres per millor visibilitat.

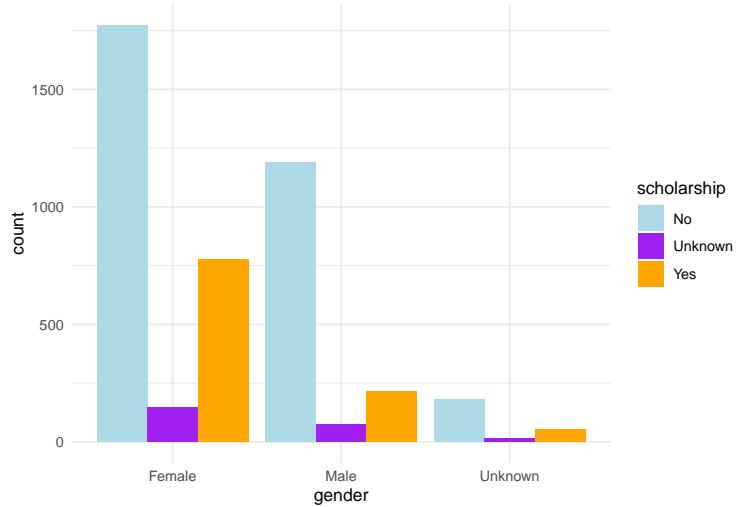


Les persones amb educació secundària com a qualificació prèvia, dos terços són dones, aproximadament.



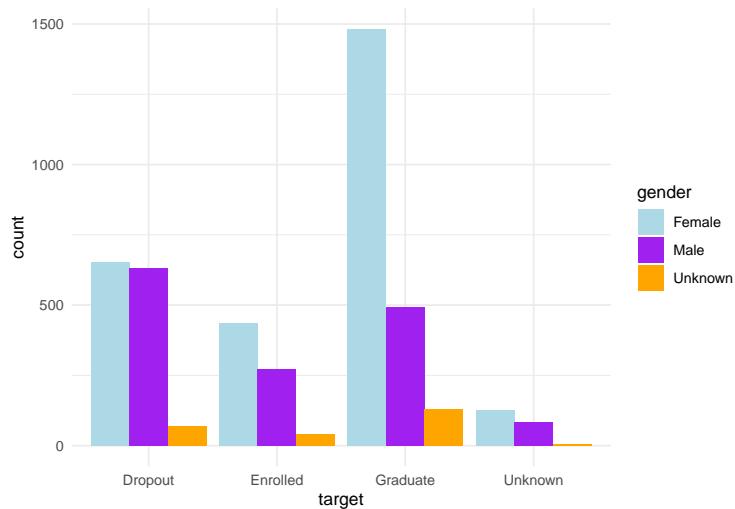
Per a les altres qualificacions prèvies, la distribució de gènere canvia. Per exemple, hi ha més persones amb un curs d'especialització tecnològica que són homes que dones, o amb educació bàsica de tercer cicle. Tot i això, cal tenir en compte que per a aquests casos la mida mostra no és gaire gran.

#### GENDER i SCHOLARSHIP



La probabilitat de tenir beca com a dona és més alta que com a home, tot i que continua sent més probable no tenir-ne.

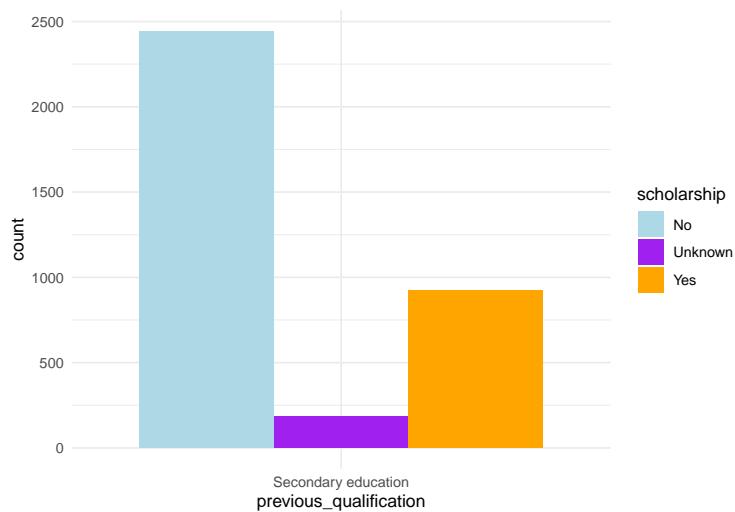
### *TARGET i GENDER*



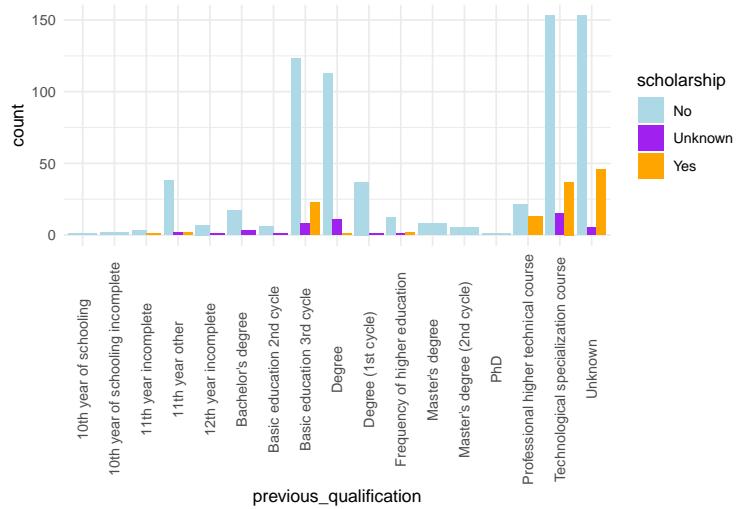
Si es miren les freqüències relatives a cada gènere, tenint en compte que hi ha més dones que homes inscrits, la probabilitat de deixar la carrera és més alta com a home, tot i que el nombre observat d'homes que l'hagin deixat i de dones sigui similar.

### *PREVIOUS\_QUALIFICATION i SCHOLARSHIP*

Per l'anàlisi de qualificació prèvia respecte de scholarship, se separaren els casos d'educació secundària respecte dels altres.

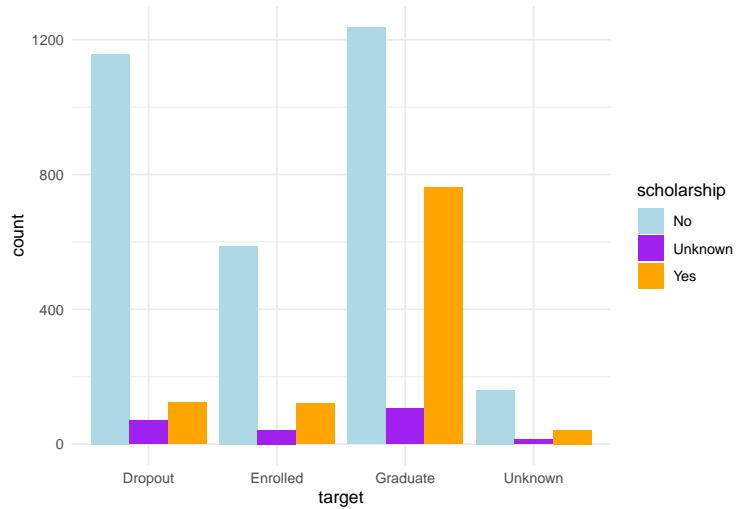


Per a persones amb educació secundària, la probabilitat de tenir beca és aproximadament un quart.



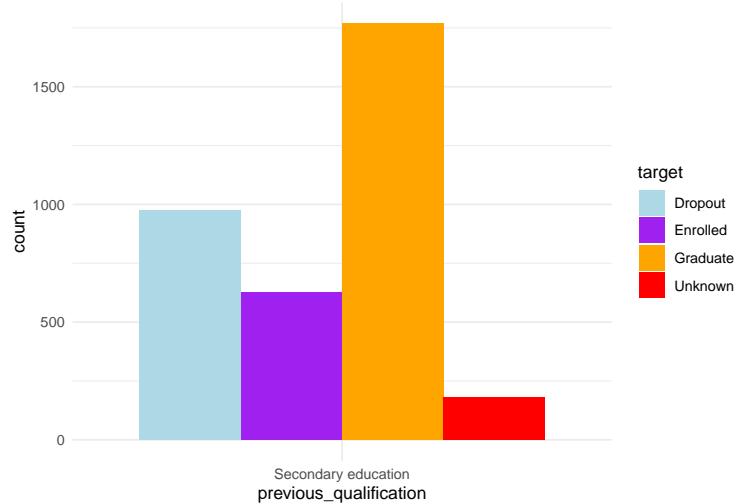
Per a la majoria dels altres casos, la probabilitat de tenir beca sembla inferior.

#### TARGET i SCHOLARSHIP

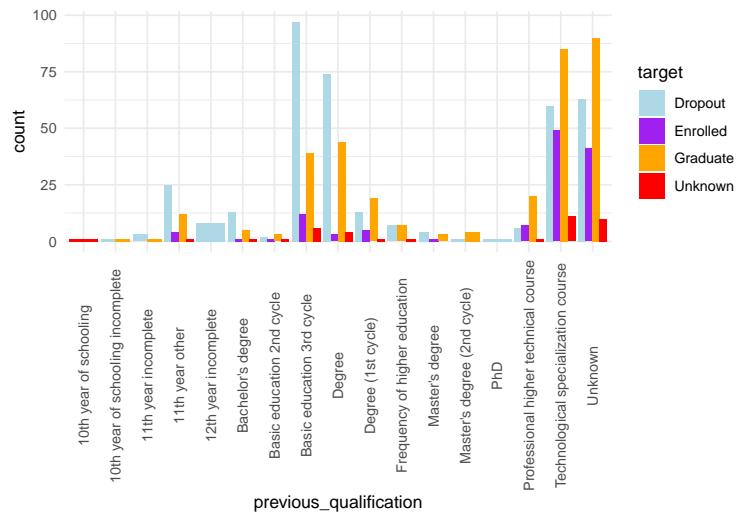


La majoria de persones amb beca s'han graduat, tot i que n'hi ha hagut que ho han deixat. Les persones sense beca tenen una probabilitat similar de deixar-ho que de graduar-se, aquesta última sent una mica més alta.

## *PREVIOUS\_QUALIFICATION i TARGET*

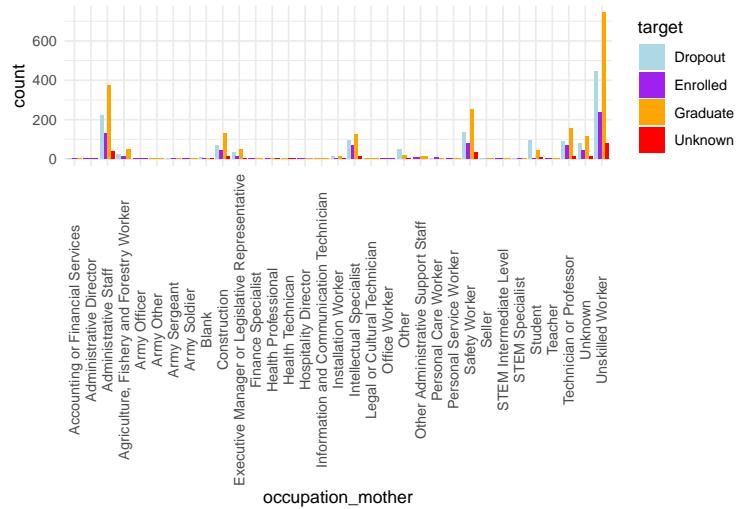


Les persones amb educació secundària tenen un terç aproximat de probabilitats de deixar-ho.



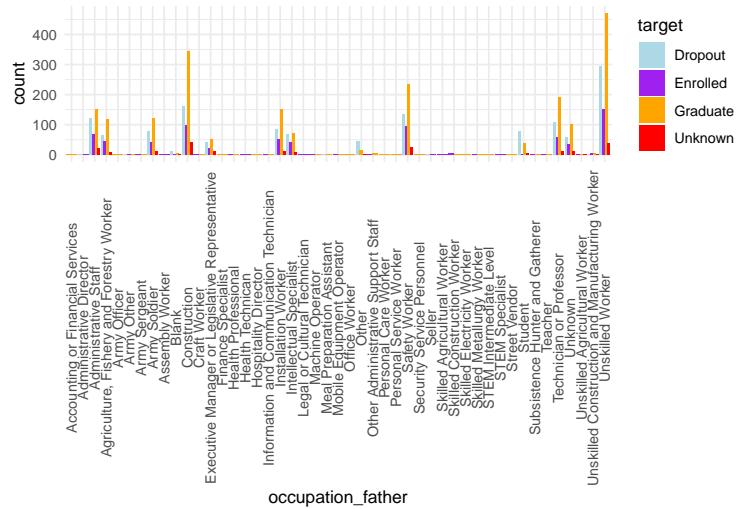
Per als altres grups la distribució és diferent. Hi ha grups amb més probabilitat de deixar-ho que de seguir, alguns de manera preocupant, com ara per a les persones amb educació bàsica de tercer cicle o de segon cicle.

### OCCUPATION\_MOTHER i TARGET



L'ocupació de la mare no sembla una variable rellevant per a si la persona es gradua o no.

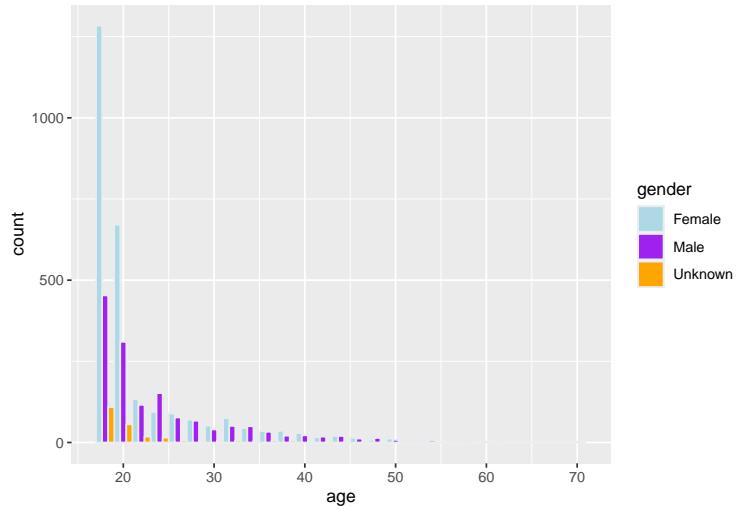
### OCCUPATION\_FATHER i TARGET



La del pare tampoc.

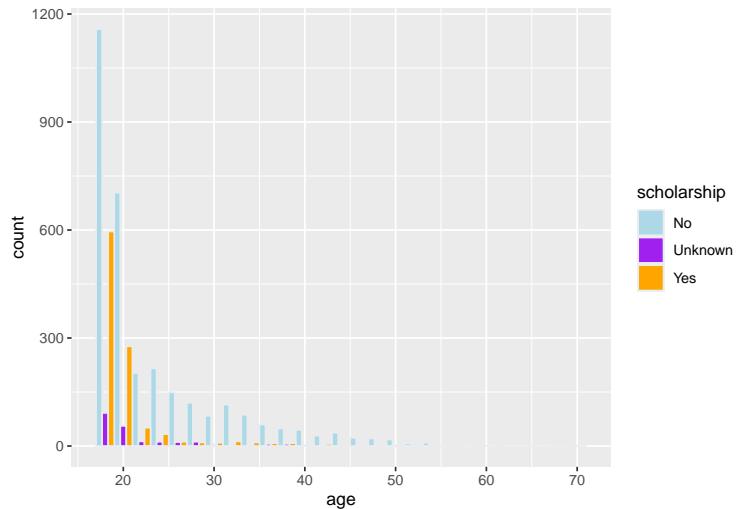
## Una variable numèrica i una de categòrica

*AGE* i *GENDER*



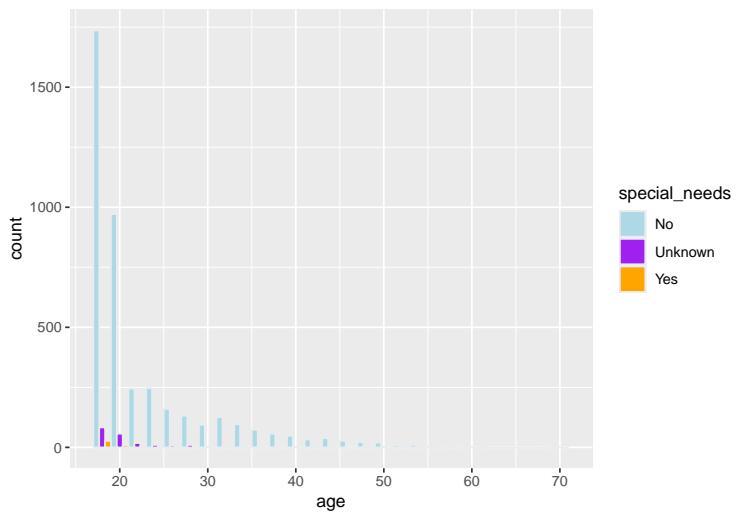
Hi ha més tendència a estudiar a edats joves. Als rangs d'edat més joves, hi ha més dones que homes. A partir, aproximadament, els 25, els dos gèneres estudien de forma similar.

*AGE* i *SCHOLARSHIP*



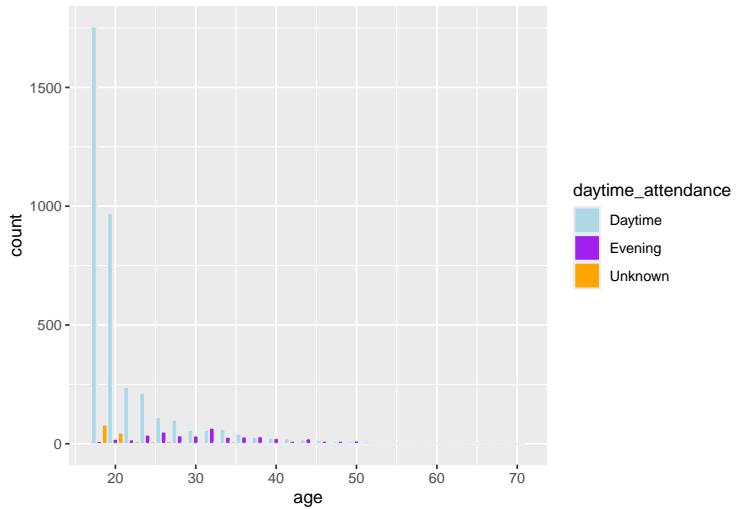
Les beques són molt més comunes a les edats més joves.

### *AGE i SPECIAL\_NEEDS*



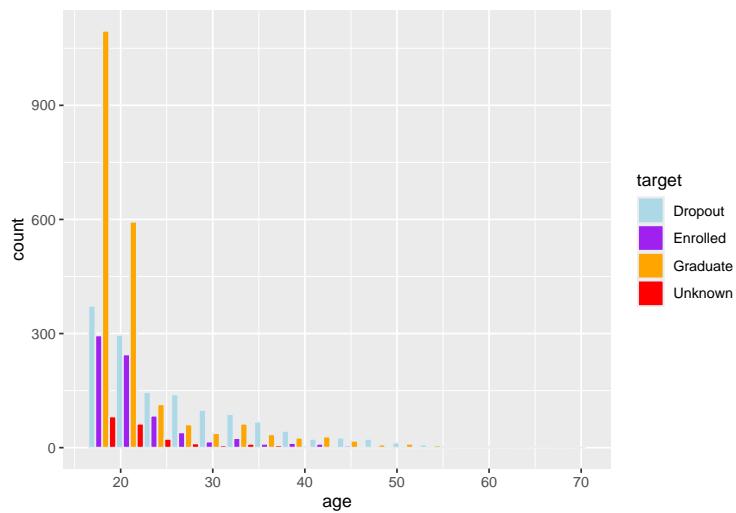
No sembla que hi hagi correlació entre l'edat i tenir necessitats especials.

### *AGE i DAYTIME\_ATTENDANCE*



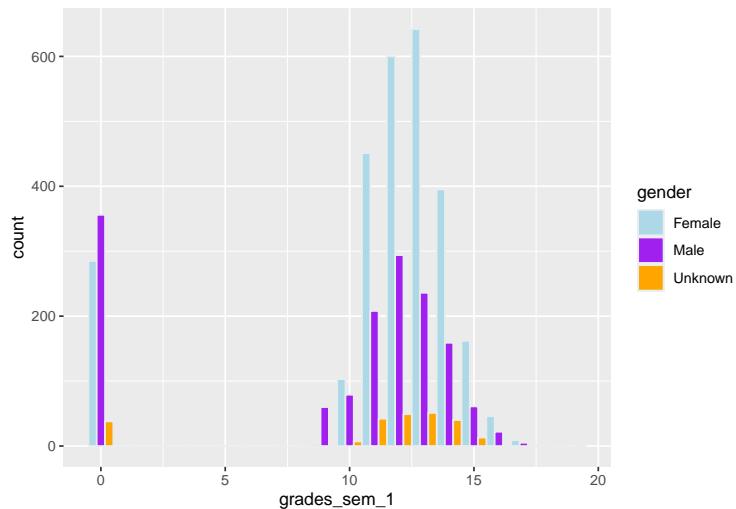
S'observa que, amb l'edat, hi ha més tendència a anar a classes de tardes que de matins.

## AGE i TARGET



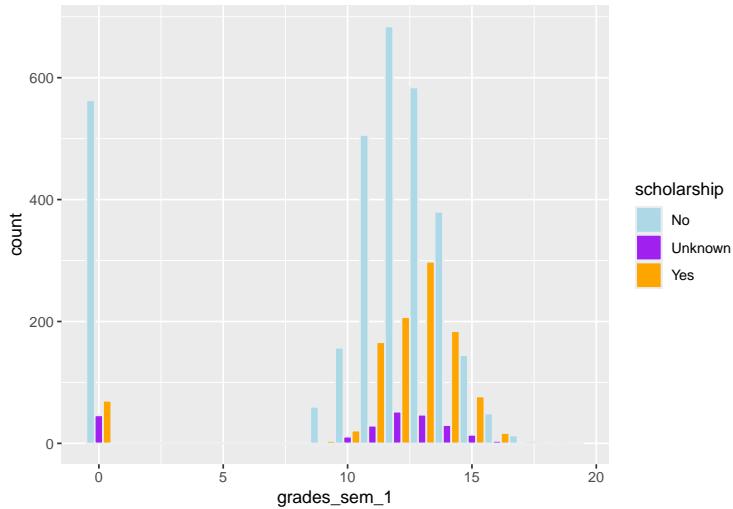
Amb l'edat, incrementa la probabilitat relativa de deixar la carrera i no graduar-se.

## GRADES\_SEM\_1 i GENDER



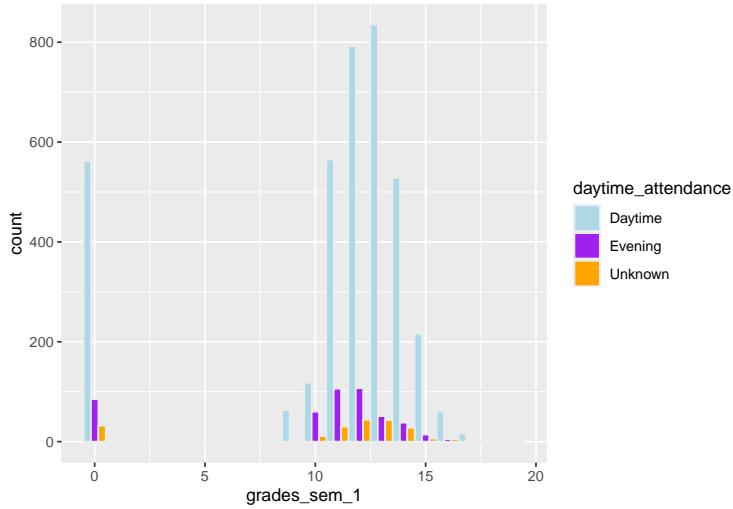
Les notes del primer semestre semblen distribuir-se igualment entre els dos gèneres, tenint les dones una mitjana relativament més alta. Tenint en compte que el 0 significa no presentat, però, s'observa que els homes tenen més tendència a no presentar-se que les dones.

### *GRADES\_SEM\_1 i SCHOLARSHIP*



Hi ha correlació entre treure notes més altes i tenir una beca. Tenint beca també és molt menys probable no presentar-se.

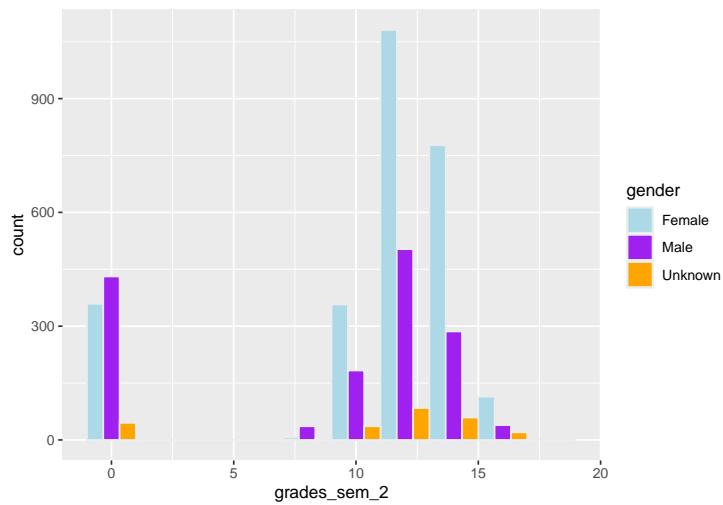
### *GRADES\_SEM\_1 i DAYTIME\_ATTENDANCE*



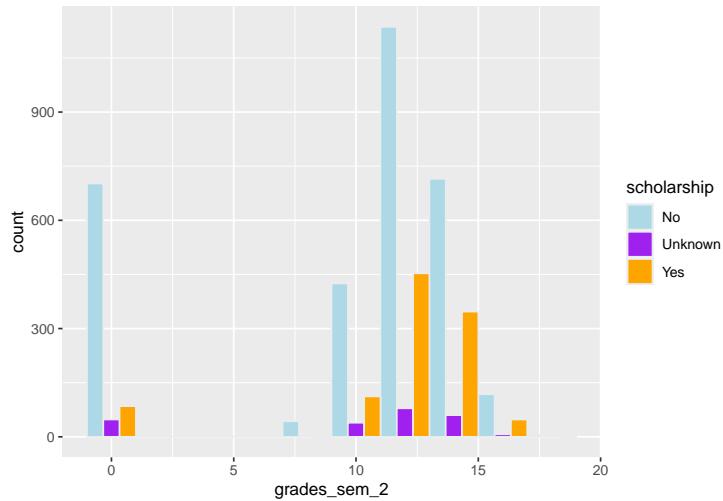
Les persones que acudeixen a classe de tardes tenen una mitjana de notes al primer semestre més baixa que les persones que acudeixen a classes al matí.

Es repliquen els mateixos patrons per a les notes del segon semestre que s'observen per a les notes del primer.

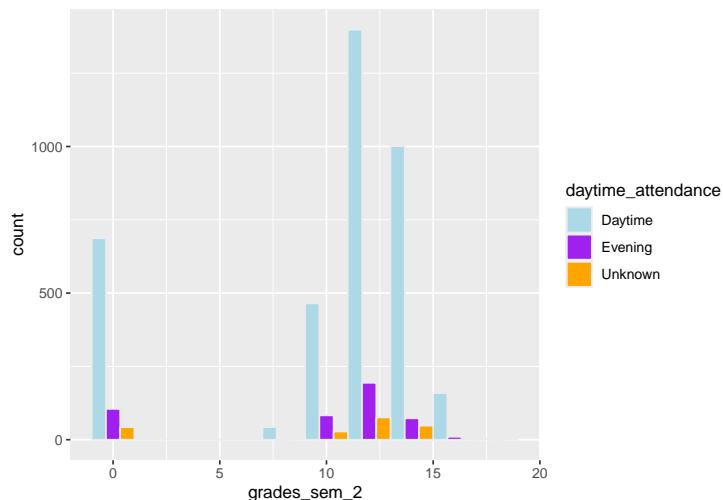
*GRADES\_SEM\_2* i GENDER



*GRADES\_SEM\_2* i SCHOLARSHIP



*GRADES\_SEM\_2* i DAYTIME\_ATTENDANCE



## 4 Clustering jeràrquic

### 4.1 Clustering

La base de dades emprada té tant variables qualitatives com quantitatives, llavors s'ha decidit analitzar-les mitjançant mètriques mixtes. Per poder fer el clustering es fa servir el mètode de Ward, que és un mètode jeràrquic ascendent. Pel mateix motiu, per computar-lo s'usa la dissimilitud de Gower al quadrat.

$$s(i, i') = \frac{\sum_{k=1}^K w_k(i, i') s_k(i, i')}{\sum_{k=1}^K w_k(i, i')}$$

$$d_{ii'}^2 = 1 - s(i, i')$$

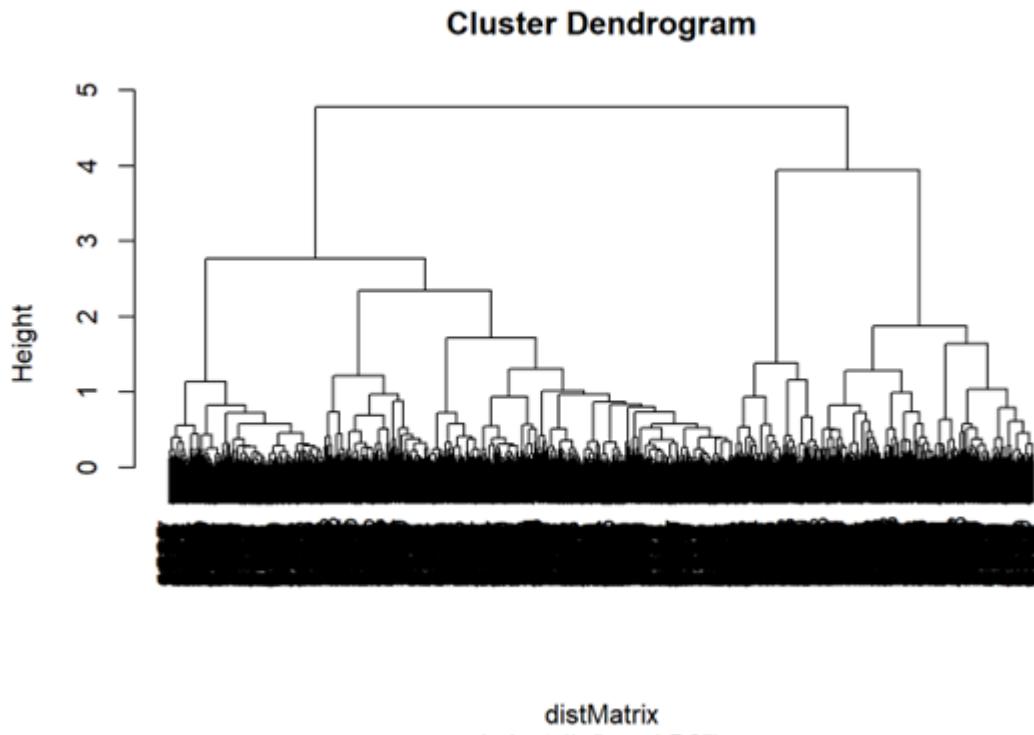
$$w_k(i, i') = \begin{cases} 0 & \text{if } (x_{ik} = NA) \text{ or if } (x_{i'k} = NA) \\ 0 & \text{if } (X_k \text{ is binary}) \text{ and } (x_{ik} = FALSE) \text{ and } (x_{i'k} = FALSE) \\ 1 & \text{otherwise} \end{cases}$$

$$s_k(i, i') = \begin{cases} 0 & \text{if } X_k \text{ numeric} \\ 0 & \text{if } (X_k = \text{qualitative}) \text{ and } (x_{ik} = (x_{i'k})) \\ 1 & \text{if } (X_k = \text{qualitative}) \text{ and } (x_{ik} \neq (x_{i'k})) \end{cases}$$

on  $R_k$  és el rang de la k-èssima variable.

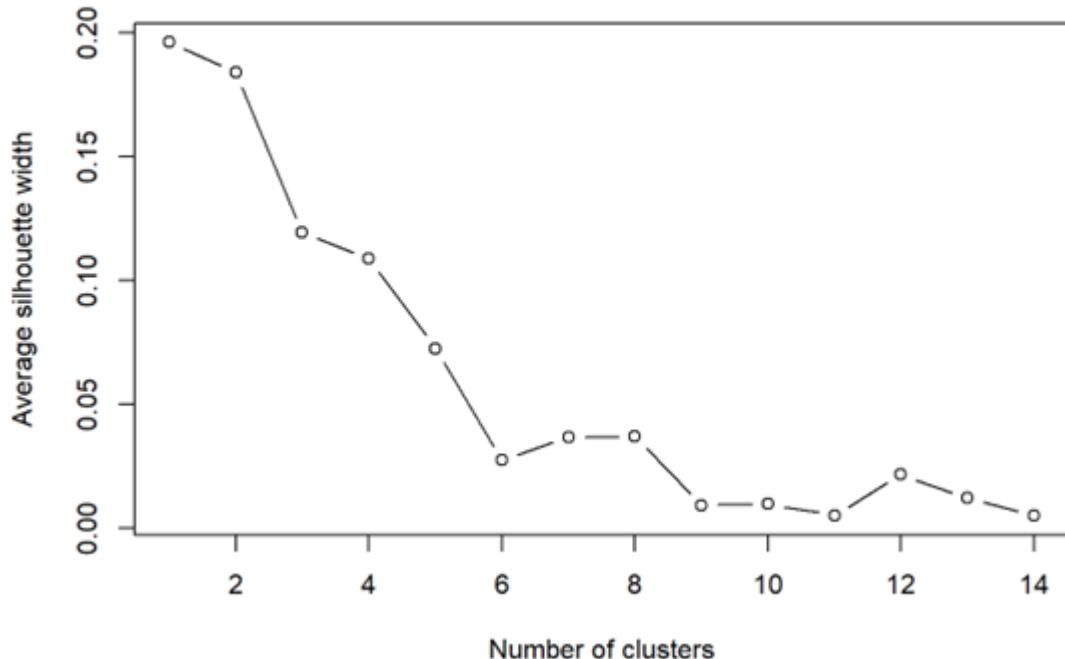
Una vegada s'ha calculat la distància, es procedeix a emprar el mètode de Ward per agrupar les dades de manera homogènia. El mètode de Ward pretén trobar a cada etapa dos clústers que proporcionen el menor increment de la suma total d'errors.

Amb la distància calculada s'aplica Ward a les dades i es fa un primer dendrograma.



En aquest dendrograma ja es veu que el nombre òptim de grups per les dades és  $k = 3$ . Encara així es comprova analíticament.

Primer, per les variables categòriques i numèriques, es fa servir el coeficient de Silhouette, que diu que  $k$  òptima és 2, encara que 3 també és una bona elecció.



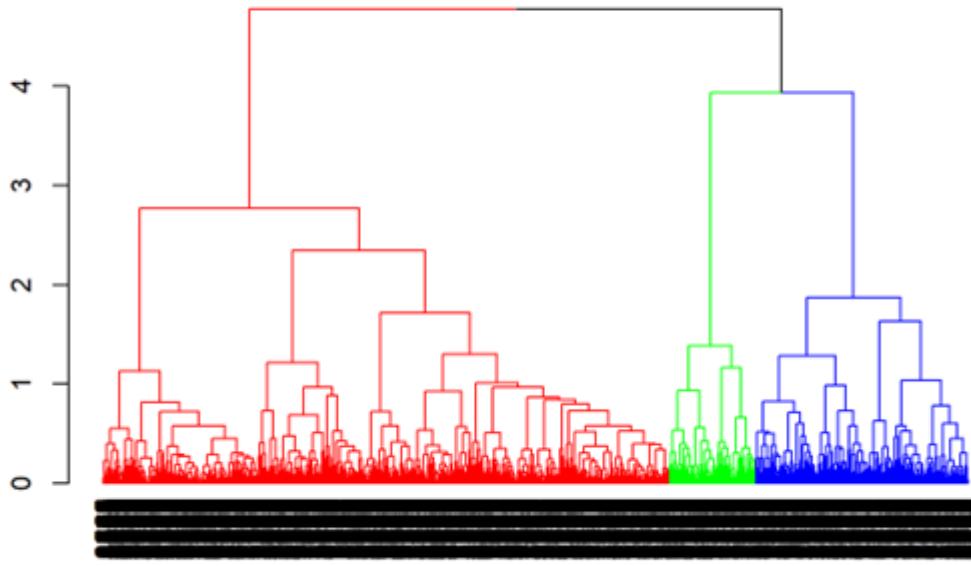
Per totes les variables numèriques, s'usa la funció de R ‘nbclust’ que indica el nombre de clústers òptim per diferents mètodes:

- Among all indices:
  - 1 proposed 2 as the best number of clusters
  - 9 proposed 3 as the best number of clusters
  - 5 proposed 4 as the best number of clusters
  - 1 proposed 5 as the best number of clusters
  - 2 proposed 7 as the best number of clusters
  - 2 proposed 10 as the best number of clusters

La majoria dels mètodes proposen  $k = 3$  com a millor partició.

Tenint en compte el primer dendrograma i aquesta informació, es decideix agrupar les dades en 3 grups diferents, per assegurar la seva homogeneïtat.

Es fa un segon clúster per visualitzar millor aquests resultats.



## 4.2 Profiling dels clusters

L'objectiu de l'elaboració de perfils és poder diferenciar entre els grups amb la distinció adequada, cosa que significa que cada grup té algunes característiques úniques sobre ell i llavors aquest cùmul és la solució ideal per segmentar els valors.

A continuació es farà una anàlisi gràfica de les variables, tant qualitatives com quantitatives, prenent com a factor els tres grups de clústers. Per tal d'observar quin d'aquests és més significatiu a les dades.

S'estudia el p-valor de les variables numèriques. S'observa quines són significatives amb el nivell de significació establert en  $\alpha = P(\text{error tipus I}) = 0.05$ .

```
##           age   grades_sem_1 no_eval_sem_1   grades_sem_2 no_eval_sem_2
## 1 3.160322e-15 0.000000e+00 6.039713e-05 0.000000e+00 8.580294e-07
## 2 0.000000e+00 1.002406e-263 3.368591e-02 1.863865e-304 9.556328e-07
## 3 8.836680e-188 1.421650e-05 4.355739e-03 1.428386e-04 2.464132e-01
##      unemployment_rate inflation_rate          gdp
## 1 1.317681e-02 2.258680e-04 1.440588e-04
## 2 2.676119e-01 7.326155e-06 1.436122e-05
## 3 1.513502e-05 3.275280e-02 7.597388e-02
```

També s'estudia el p-valor de les variables categòriques per triar les significatives, mitjançant el test chi-quadrat de cadascuna respecte als clústers.

```

##          course      daytime_attendance previous_qualification
## 0.000000e+00          0.000000e+00          5.097703e-56
## nationality      occupation_mother      occupation_father
## 9.863178e-01          6.059069e-37          2.946157e-21
## special_needs           gender      scholarship
## 1.858605e-01          1.558903e-61          2.392317e-63
## target
## 0.000000e+00

```

A la següent taula es mostra el valor 1 de les variables considerades com a significatives. Aquestes seran estudiades gràficament com prèviament s'ha esmentat.

```

##          course      daytime_attendance previous_qualification
## 1              1                  1                  1
## nationality      occupation_mother      occupation_father
## 0              0                  1                  1
## special_needs           gender      scholarship
## 0              0                  1                  1
## age            age      grades_sem_1      no_eval_sem_1
## 1              1                  1                  1
## grades_sem_2      no_eval_sem_2      unemployment_rate
## 1              1                  0                  0
## inflation_rate           gdp      target
## 1              1                  1                  1

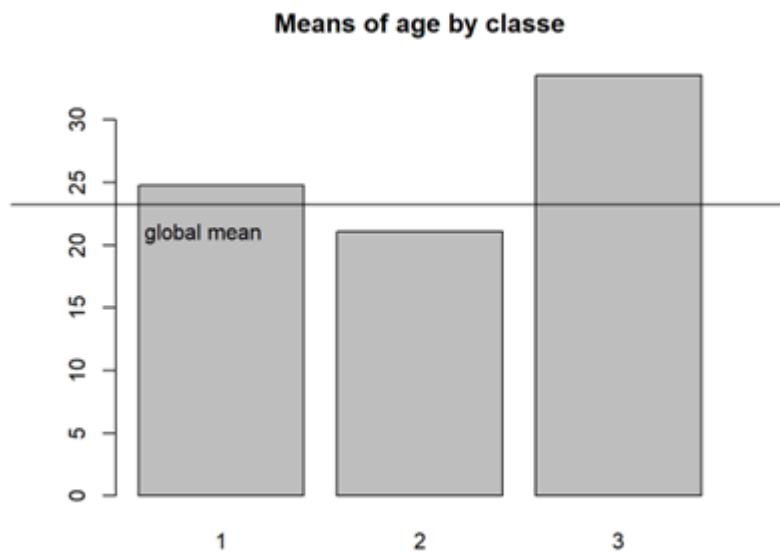
```

S'exclouen de l'anàlisi les següents variables:

- Nationality
- Special\_needs
- No\_eval\_sem\_2
- Unemployment\_rate

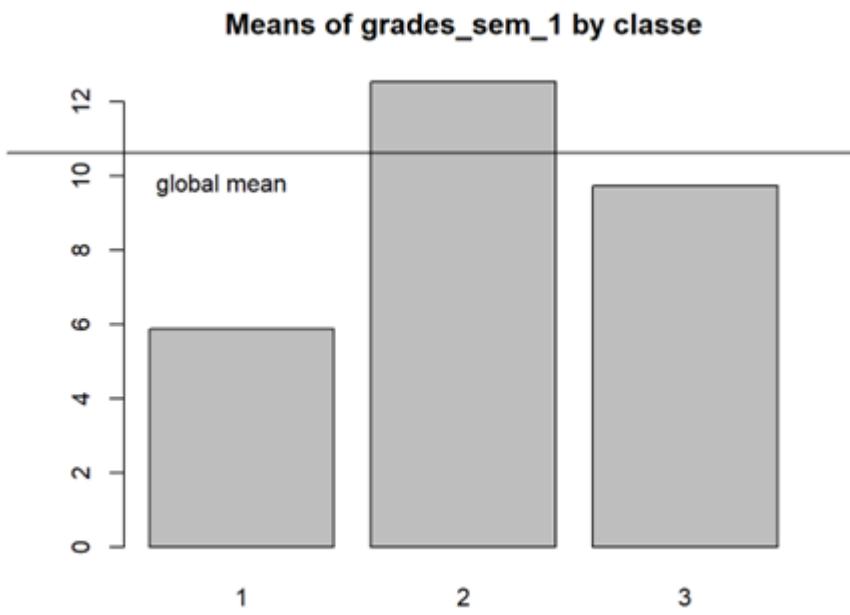
#### 4.2.1 Anàlisi gràfica de les variables numèriques

*EDAT*



A aquest gràfic es veu com els clústers 1 i 2 són els més propers a la mitjana, mentre el clúster 3, s'allunya més, el que indica que és el més significatiu per a la variable edat.

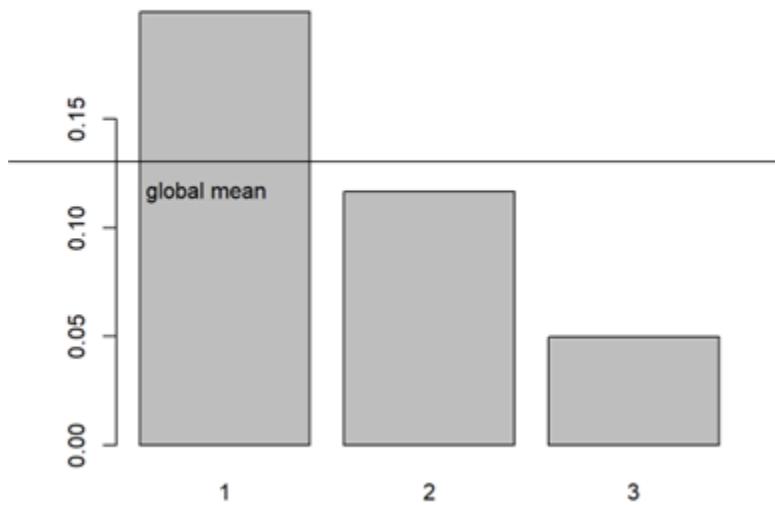
*NOTES AL PRIMER SEMESTRE*



Per a la variable que tracta sobre les notes al primer semestre, el clúster més significatiu és el corresponent al grup 1.

*UNITATS CURRICULARS NO AVALUADES AL PRIMER SEMESTRE*

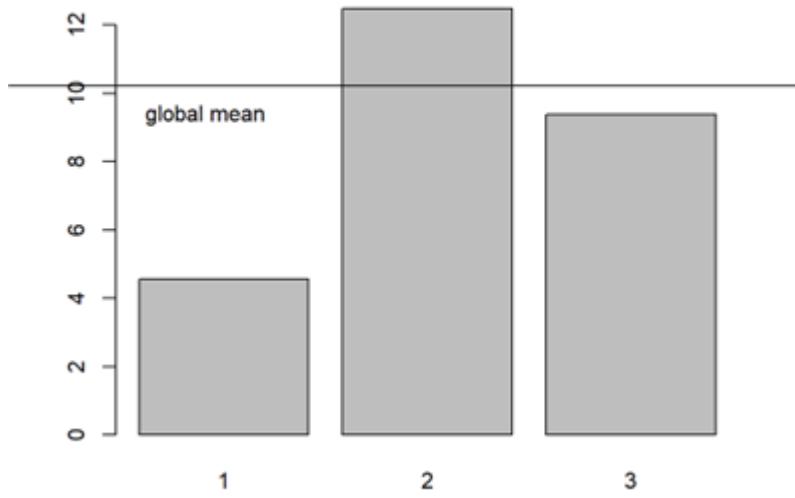
**Means of no\_eval\_sem\_1 by classe**



A aquest gràfic es veu com els tres grups varien molt de la mitjana, és a dir, que tots tres són significatius a aquesta variable.

*NOTES AL SEGON SEMESTRE*

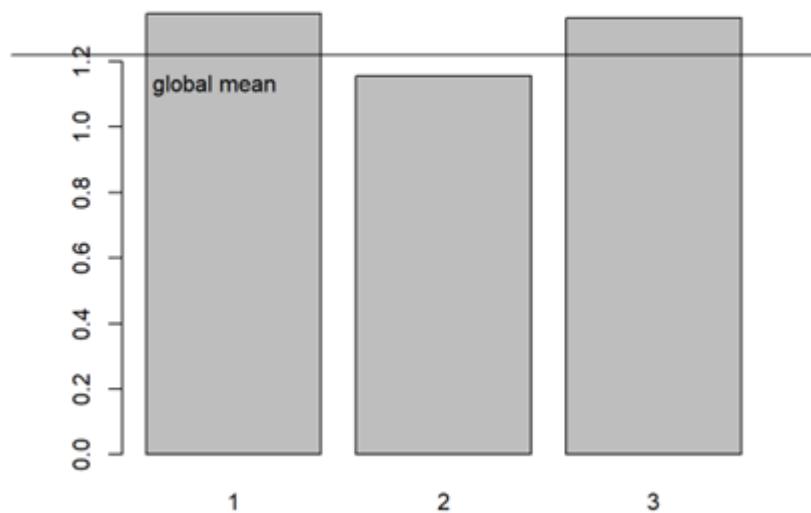
**Means of grades\_sem\_2 by classe**



Pel que fa als grups en les qualificacions del segon semestre, coincideix amb el gràfic obtingut en les qualificacions del primer. Per tant, el grup significatiu per a aquest grup és el primer.

## TAXA D'INFACIÓ

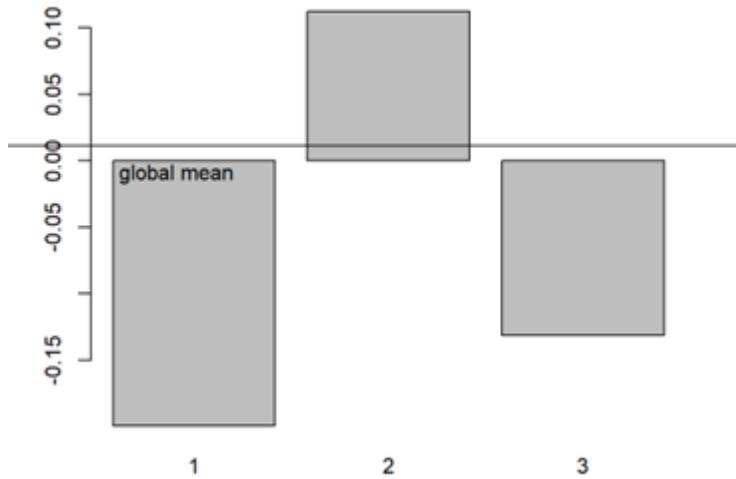
**Means of inflation\_rate by classe**



Els tres clústers es troben quasi a la mateixa altura respecte a la mitjana, són molt semblants, no expliquen la variable significativament.

## PIB

**Means of gdp by classe**



Observant els valors que prenen, el clúster 2 és el més allunyat de la mitjana, per tant, és el més significatiu. Els altres són molt semblants.

Com a conclusió general, sobre els clústers a les variables numèriques, el clúster 1 és el més significatiu, sent el millor a la meitat de les variables estudiades. A l'altra meitat de les variables els clústers 2 i 3 també destaquen a una variable cadascun.

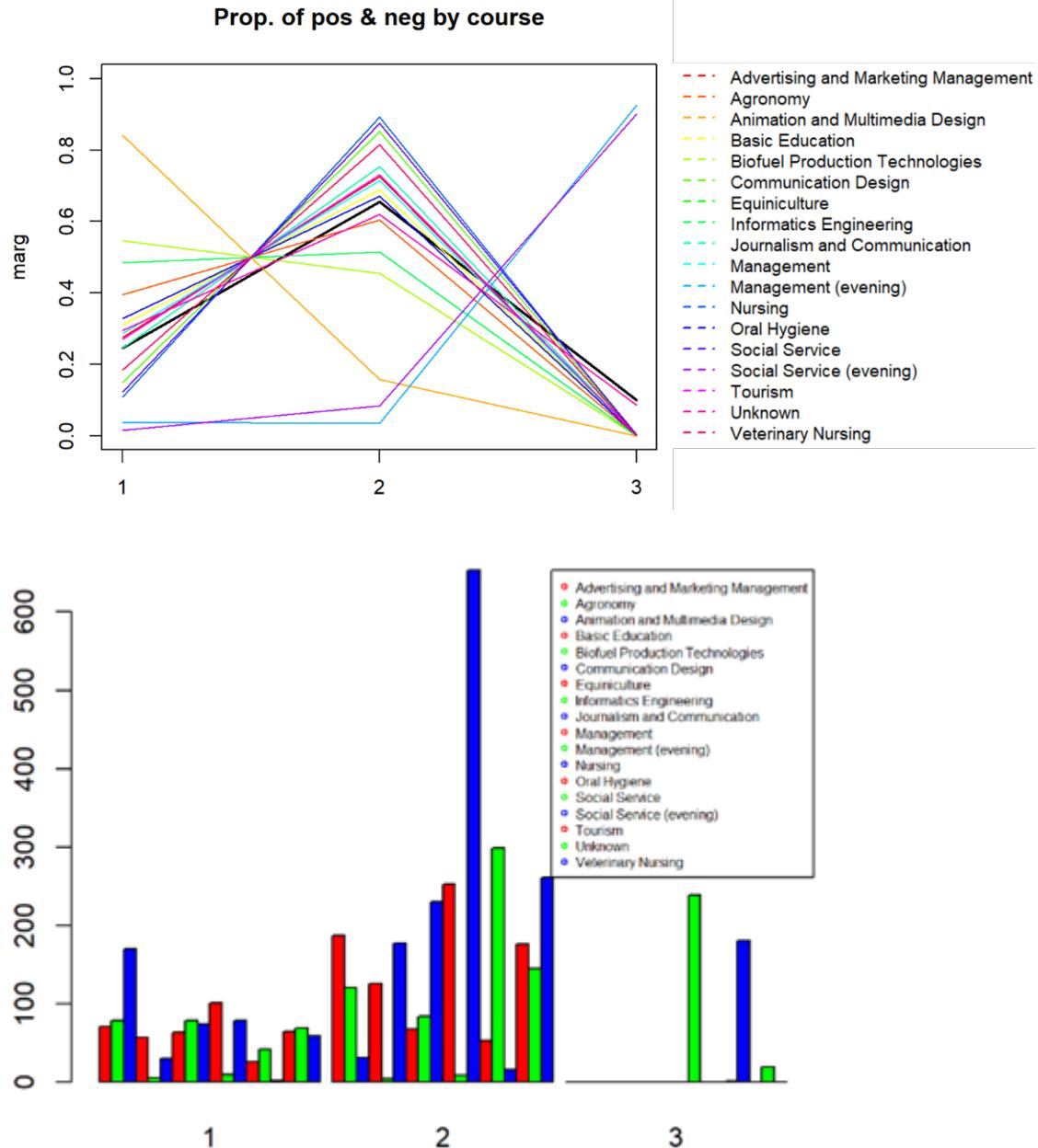
**Clúster 1:** Destaca en les notes més baixes per als dos semestres i la quantitat més gran d'unitats curriculars suspeses.

**Clúster 2:** Destaca a les notes més altes als dos semestres i el PIB més gran.

**Clúster 3:** Destaca en les edats més grans i el menor nombre d'unitats curriculars suspeses.

#### 4.2.2 Anàlisi gràfica de les variables categòriques

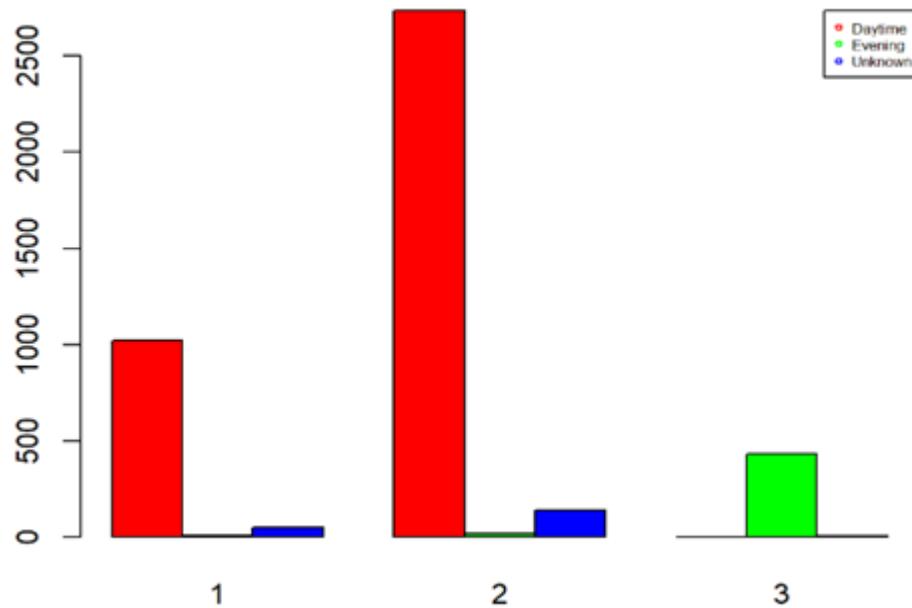
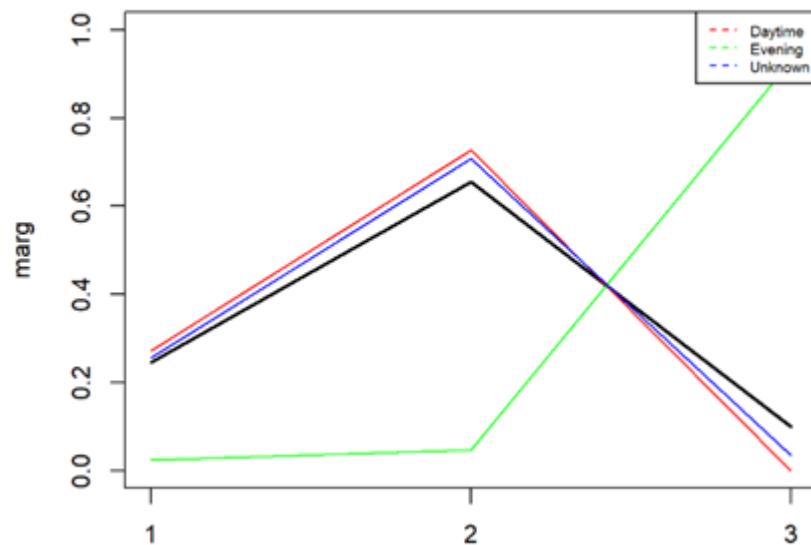
*GRAU DE L'ESTUDIANT*



El primer gràfic mostra com la majoria de les categories a la variable es distribueixen als grups. Cal destacar la divergència “Animation Multimedia Design”, “Social service” i “Management”, aquestes dues últimes categories sent la modalitat de tarda. Al segon gràfic es veu que el grup 2 explica més quantitat de les dades que no pas el grup 3, que quasi no explica.

## HORARI DE L'ESTUDIANT

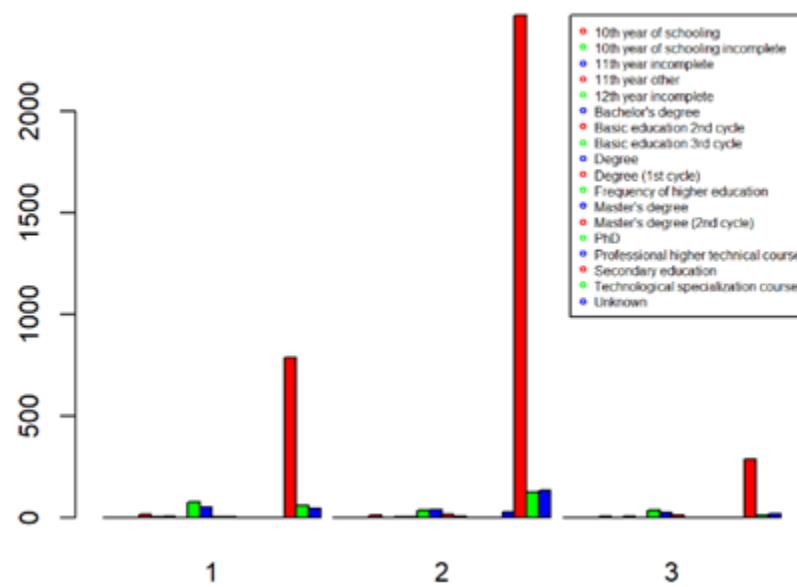
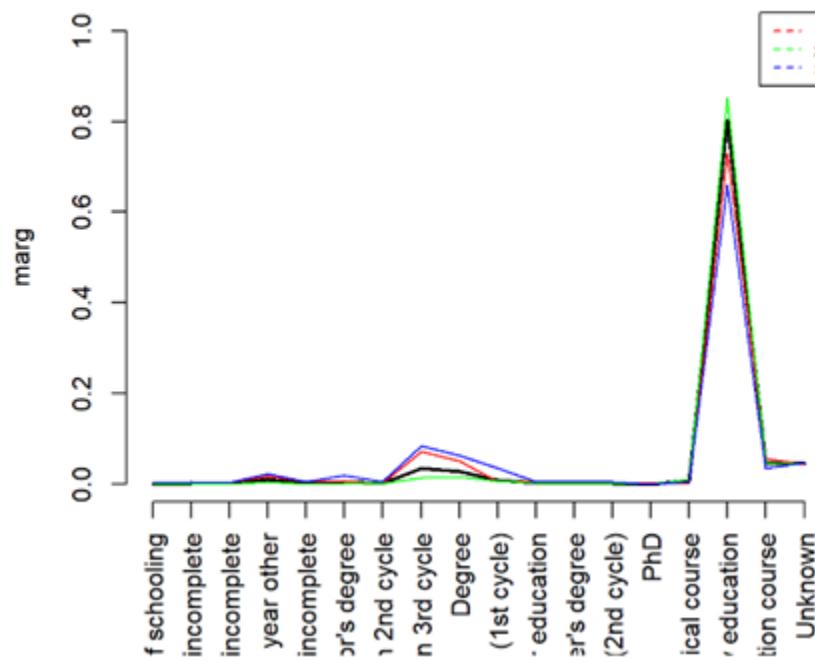
Prop. of pos & neg by daytime\_attendance



Als dos gràfics s'observa com els grups 1 i 2 són els que expliquen els estudiants de matins, especialment el grup 2. El grup 3 és el que explica els estudiants de tardes.

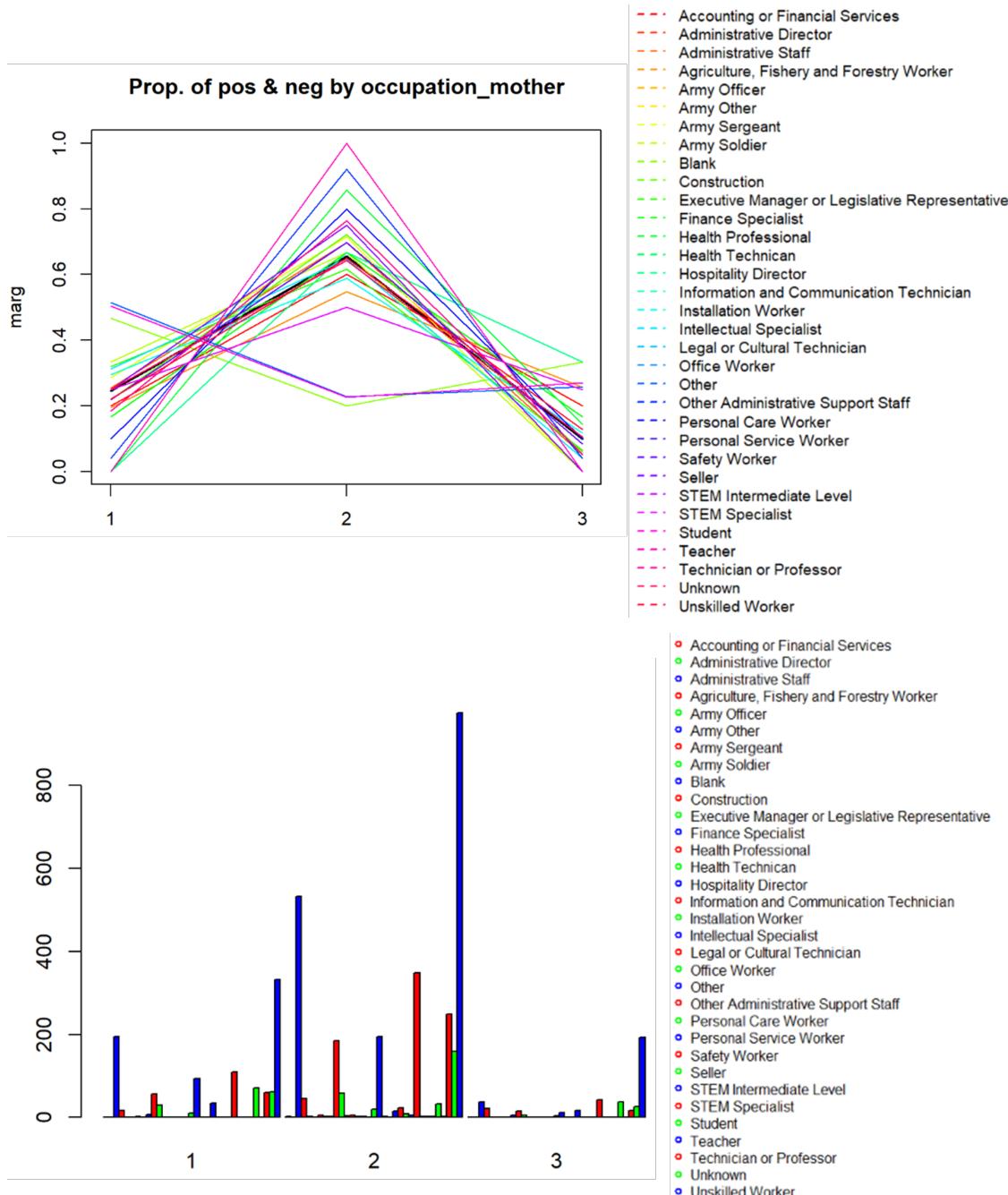
*QUALIFICACIÓ PRÈVIA DE L'ESTUDIANT*

**Prop. of pos & neg by previous\_qualification**



S'observa com els estudiants amb un títol previ de secundària són els que més es mostren en la proporció, seguint la línia de la proporció global als grups, representada amb color negre al primer gràfic.

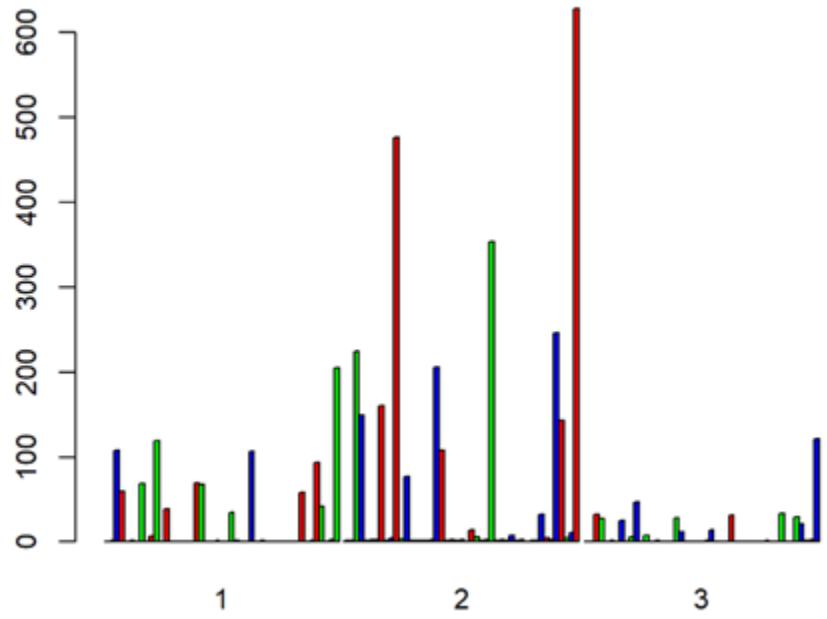
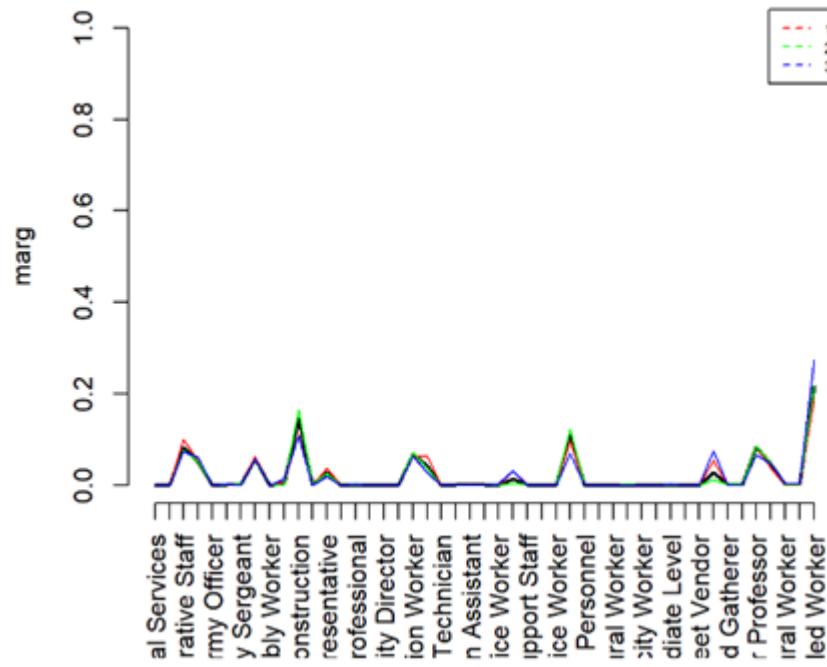
## Ocupació de la mare



Pel que fa a les categories sobre el treball de les mares, totes les ocupacions segueixen la mateixa proporció als grups menys tres, "Other", "Bank" i "Student". Al segon gràfic s'observa que: "Unskilled worker" és la categoria amb més proporció entre totes. El grup dos es caracteritza per tenir el nombre més gran de mares que són professors.

## Ocupació del Pare

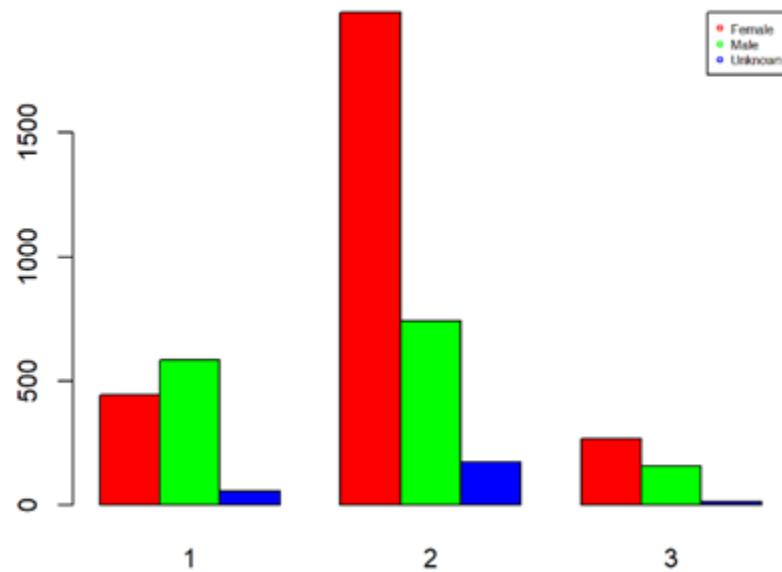
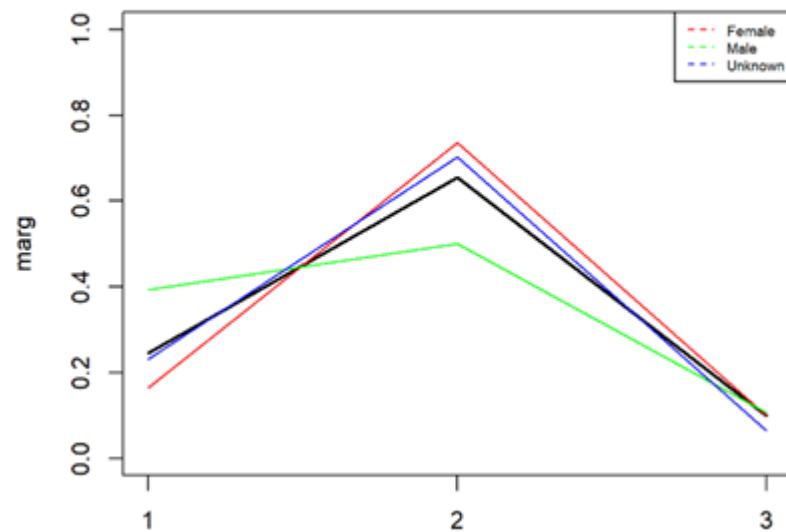
Prop. of pos & neg by occupation\_father



Per altra banda, a l'ocupació del pare, les categories es distribueixen en els grups segons la línia mitjana. Proporcionalment, el segon grup explica més categories, seguint per les restants al grup 1 i quantitats petites al 3.

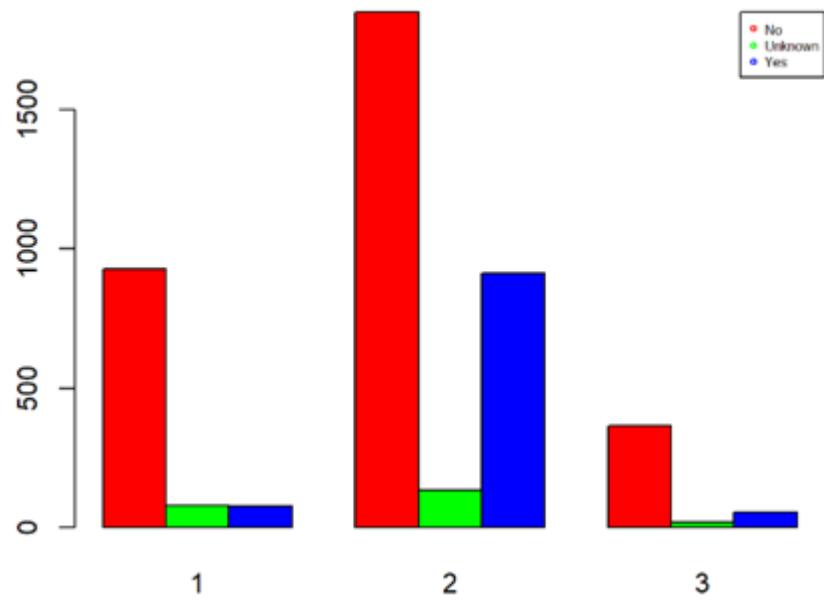
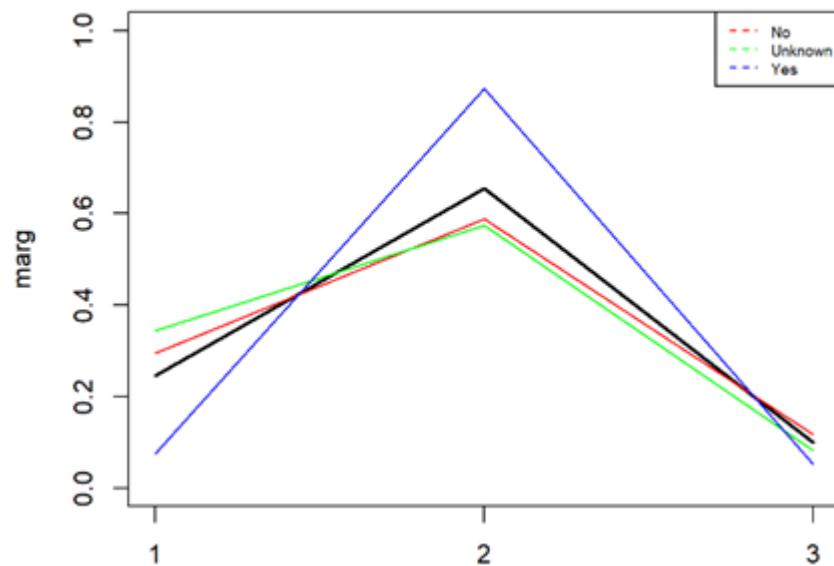
## GÈNERE

Prop. of pos & neg by gender



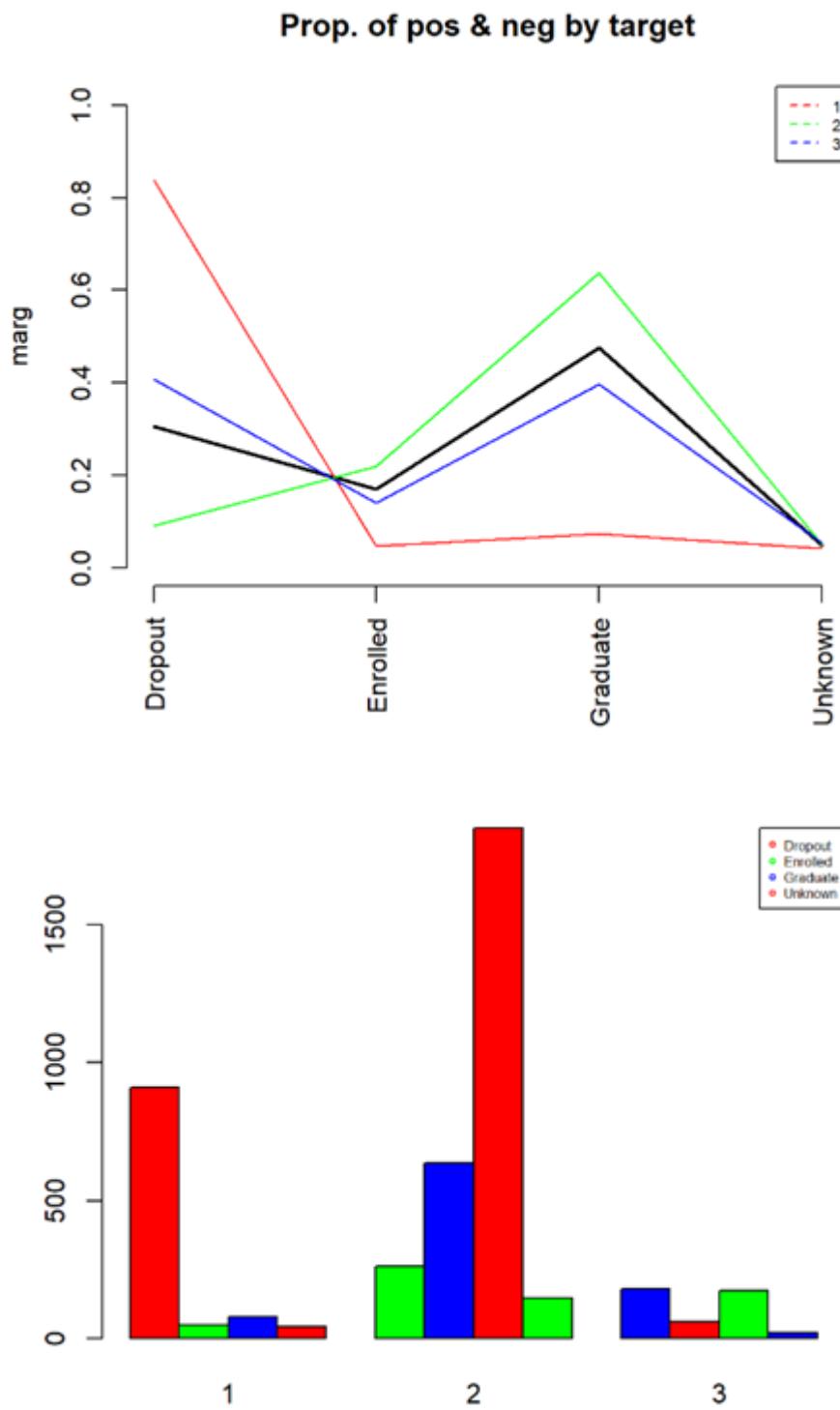
Tractant el gènere dels estudiants, el primer grup acull a quasi el mateix nombre d'homes i dones, igual que el tercer encara que aquest menys. El segon grup destaca en la seva gran proporció de dones.

### Prop. of pos & neg by scholarship



Al primer gràfic es veu com totes les categories es distribueixen de la mateixa manera. Al segon gràfic s'observa l'alta proporció d'estudiants amb beca.

*RESULTAT ACADÈMIC*



Als dos gràfics s'observa com el primer grup està constituït pels estudiants que han deixat d'estudiar. El segon, que és el més gran en proporció format pels graduats i els que continuen estudiant.

#### **4.2.3 Conclusió sobre els clústers a les variables numèriques i categòriques**

**Clúster 1:** Destaca en les notes més baixes per als dos semestres i la quantitat més gran d'unitats curriculars suspeses. És el grup dels estudiants que deixen els estudis i estudiants sense beca.

**Clúster 2:** Destaca a les notes més altes als dos semestres i el PIB més gran. És el grup dels estudiants graduats, amb beca, majoritàriament dones i estudiants de matins, a més de tenir el nombre més gran de mares professors.

**Clúster 3:** Destaca en les edats més grans i el menor nombre d'unitats curriculars suspeses. És el grup format per estudiants amb horari de tarda. En la mateixa proporció d'estudiants que es graduen, com els que deixen d'estudiar.

## 5 ACP

S'observa que la base de dades té un total de 8 columnes numèriques. Per tant, l'anàlisi de components principals tindrà 8 components.

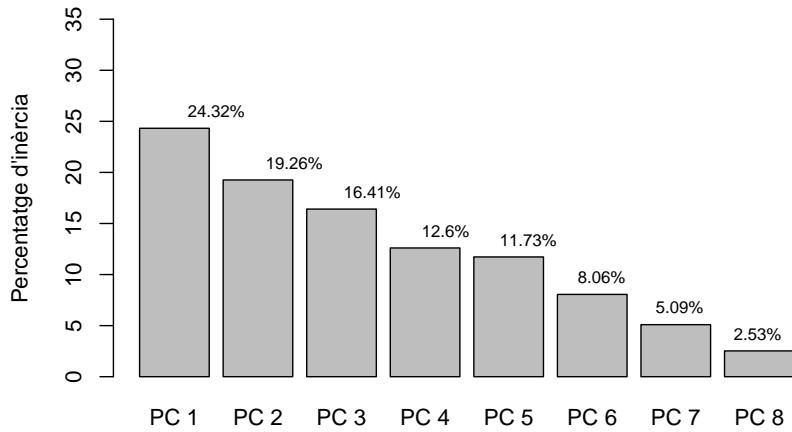
```
##           age      grades_sem_1      no_eval_sem_1      grades_sem_2
## "numeric"    "numeric"    "numeric"    "numeric"
## no_eval_sem_2 unemployment_rate   inflation_rate      gdp
## "numeric"    "numeric"    "numeric"    "numeric"
```

Els resultats de l'anàlisi de components principals sobre les variables numèriques són els següents:

```
## Standard deviations (1, ..., p=8):
## [1] 1.3948881 1.2413107 1.1457856 1.0041560 0.9685334 0.8028912 0.6383884
## [8] 0.4494999
##
## Rotation (n x k) = (8 x 8):
##          PC1       PC2       PC3       PC4       PC5
## age      0.24586659  0.03529476 -0.04350367  0.03379747 -0.96650110
## grades_sem_1 -0.59978174 -0.32227631 -0.07909503  0.05747587 -0.17266379
## no_eval_sem_1  0.30271755 -0.61710448  0.13456221  0.05090417  0.05602916
## grades_sem_2   -0.60281108 -0.32703352 -0.04403668  0.05853729 -0.14805220
## no_eval_sem_2   0.30585628 -0.60477416  0.14291011  0.04943162  0.05057019
## unemployment_rate 0.04184307 -0.05563562 -0.68285536 -0.33822281  0.04639942
## inflation_rate   0.03680361  0.09645898 -0.20770896  0.93197266  0.06274800
## gdp            -0.16774671  0.17021150  0.66479128 -0.06422781 -0.04854441
##          PC6       PC7       PC8
## age      0.026553270  0.009832622  0.018518454
## grades_sem_1 -0.007957304 -0.016242952 -0.704769753
## no_eval_sem_1 -0.038854424  0.708477929 -0.016000326
## grades_sem_2   -0.014794349  0.003771702  0.708625067
## no_eval_sem_2   0.231872481 -0.679328477  0.004898898
## unemployment_rate 0.630751286  0.118979092  0.017652568
## inflation_rate   0.266673044  0.050541810  0.004337384
## gdp            0.689040290  0.139594434 -0.014026455
```

Cada component representa una inèrcia concreta. S'explica el percentatge d'inèrcia que representa cada component.

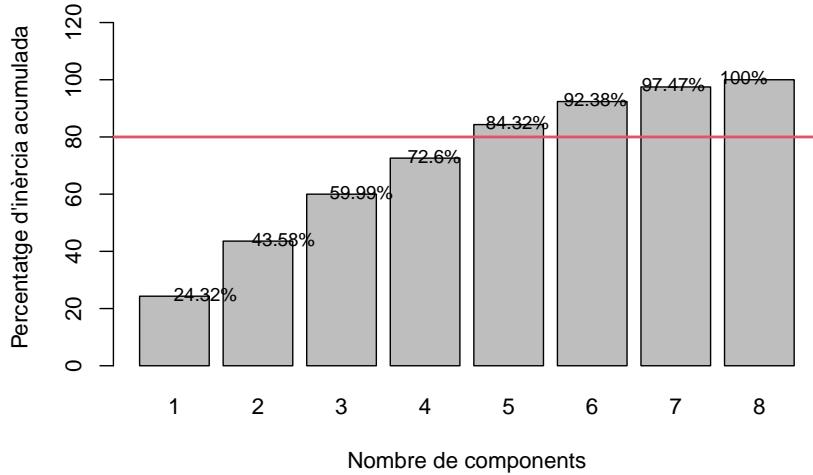
### Percentatge del total d'inèrcia en cada component principal



Es veu que la primera component representa el 24.32% de la inèrcia, la segona el 19.26%, la tercera el 16.41%, la quarta el 12.6%, la cinquena el 11.73%, la sisena el 8.06%, la setena el 5.09% i, la vuitena el 2.53%.

Interessa també la inèrcia acumulada:

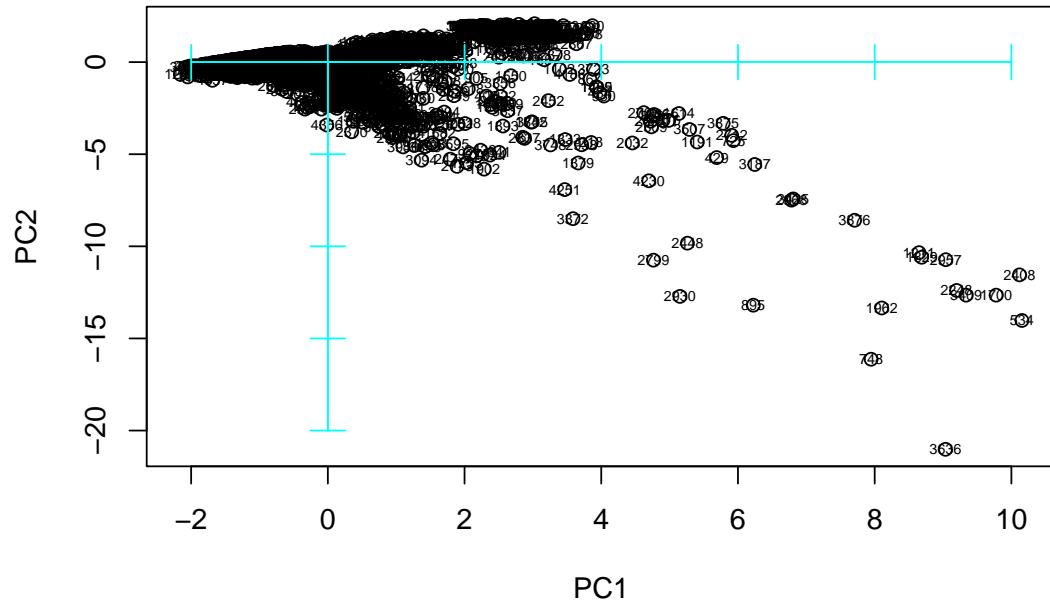
### Percentatge d'inèrcia acumulada en cada Component Principal



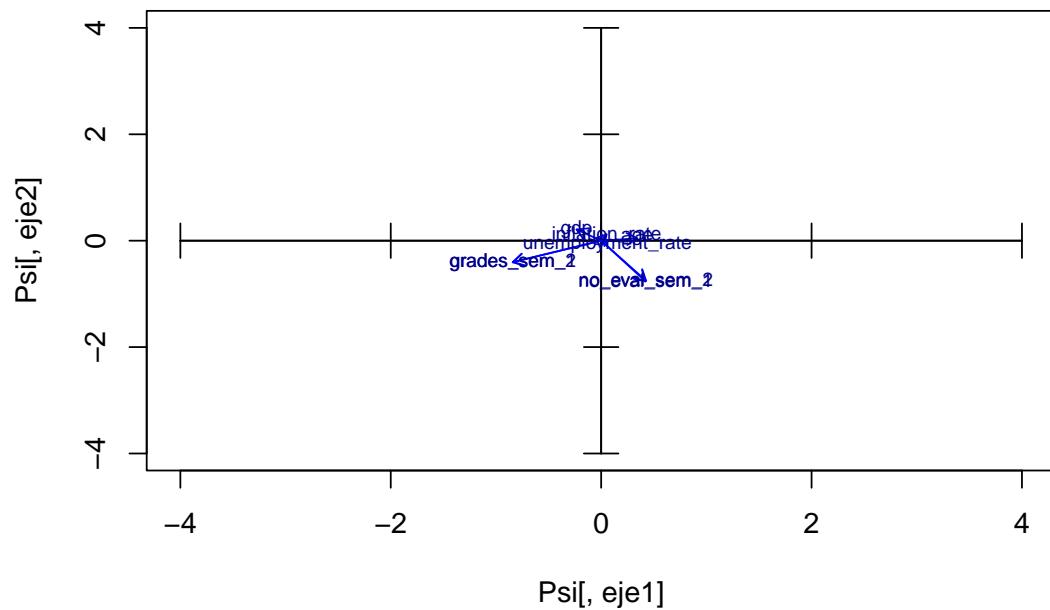
Tenint en compte que la inèrcia equival a la proporció de la variabilitat de les dades, se sap que amb un 80% d'inèrcia, es pot obtenir gairebé tota la informació o variabilitat de la base de dades original. En el gràfic de la inèrcia acumulada es pot veure que amb les 4 primeres components ja s'aconsegueix gairebé el 80% de la inèrcia. Si es pren les 5 primeres components, el total de variabilitat que s'aconsegueix explicar és del 84.32%. Ens quedem amb les components PC1, PC2, PC3, PC4 i PC5.

Es realitza un gràfic de dispersió per a totes les combinacions possibles:

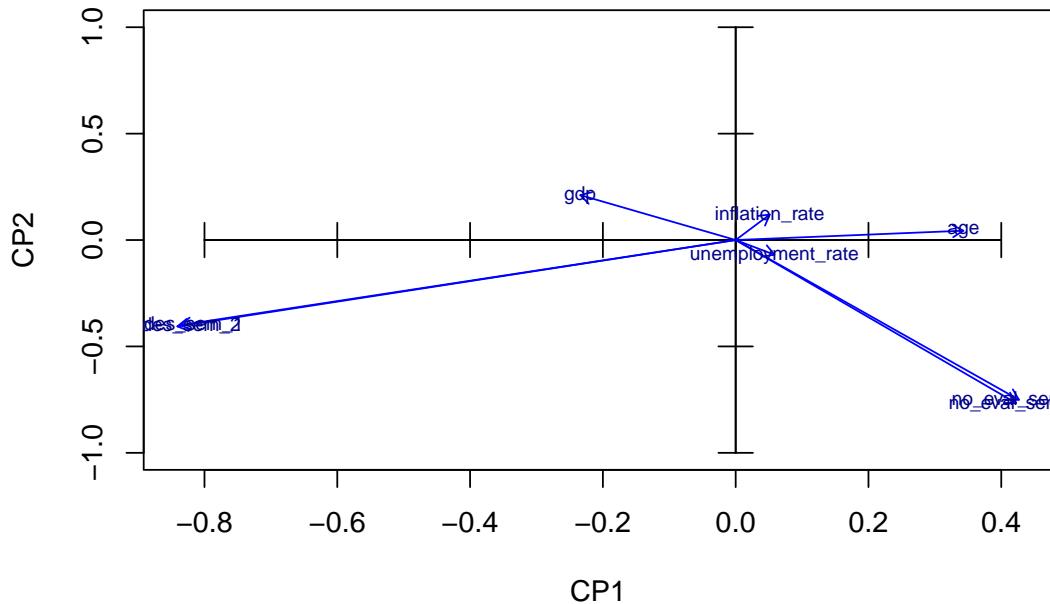
## Projecció d'observacions entre PC1 i PC2



Les observacions es concentren principalment en els valors més petits de les dues components.

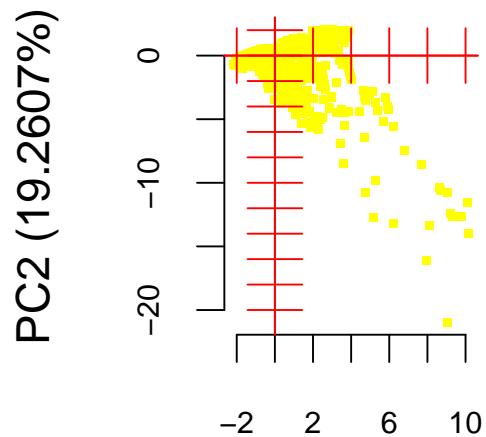


## 5.1 Projecció de les variables numèriques

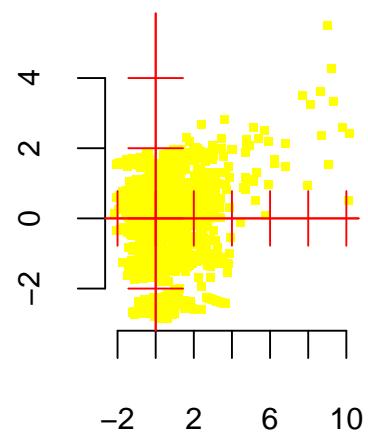


Les variables que representen les notes dels semestres i les variables que fan referència a les unitats no avaluables són les que tenen més variabilitat continguda en la representació dels dues components del biplot. Així mateix, s'observa que l'edat ha contribuït bastant a la component 1, ja que es veu com el vector paral·lel a l'eix de la component 1.

Gràfic de dispersió dels individus



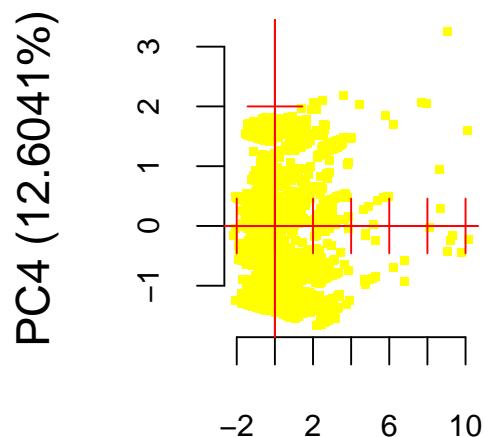
Gràfic de dispersió dels individus



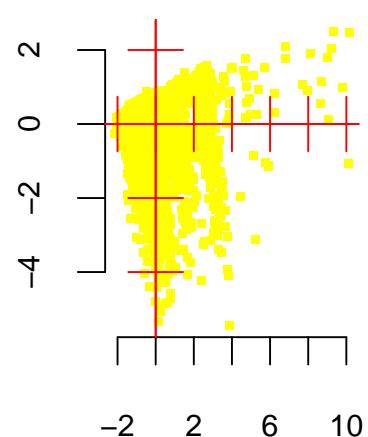
PC1 (24.3214%)

PC1 (24.3214%)

Gràfic de dispersió dels individus



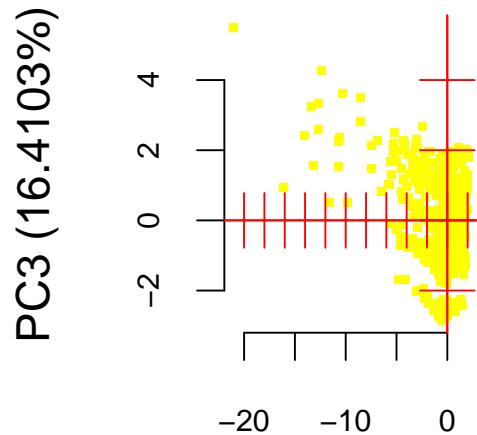
Gràfic de dispersió dels individus



PC1 (24.3214%)

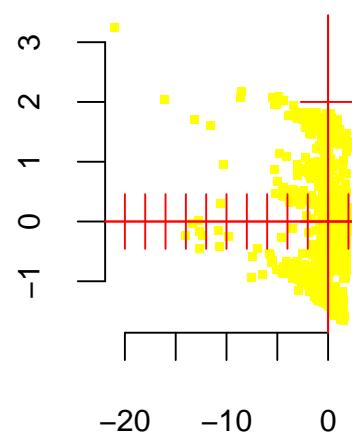
PC1 (24.3214%)

Gràfic de dispersió dels individus



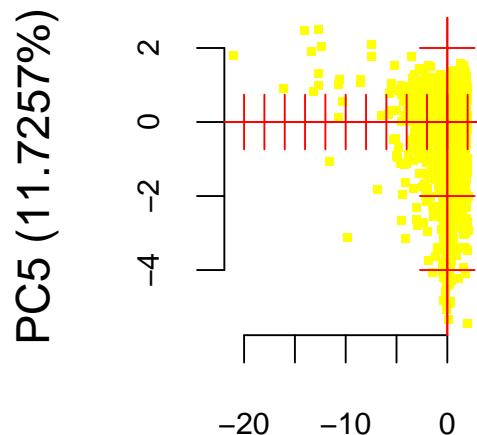
PC3 (16.4103%)

Gràfic de dispersió dels individus



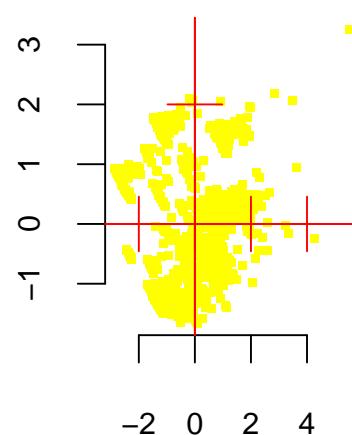
PC2 (19.2607%)

Gràfic de dispersió dels individus



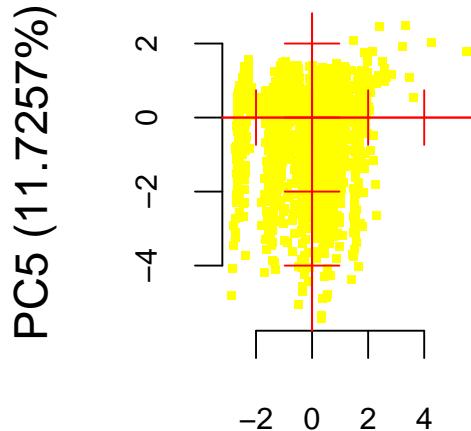
PC2 (19.2607%)

Gràfic de dispersió dels individus



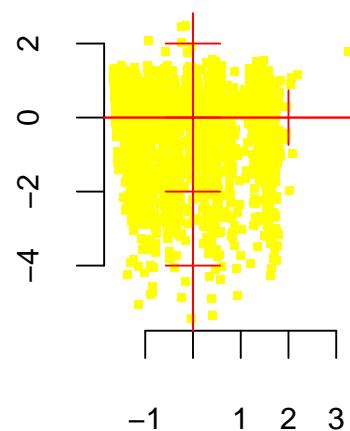
PC3 (16.4103%)

**Gràfic de dispersió dels individus**



**PC3 (16.4103%)**

**Gràfic de dispersió dels individus**



**PC4 (12.6041%)**

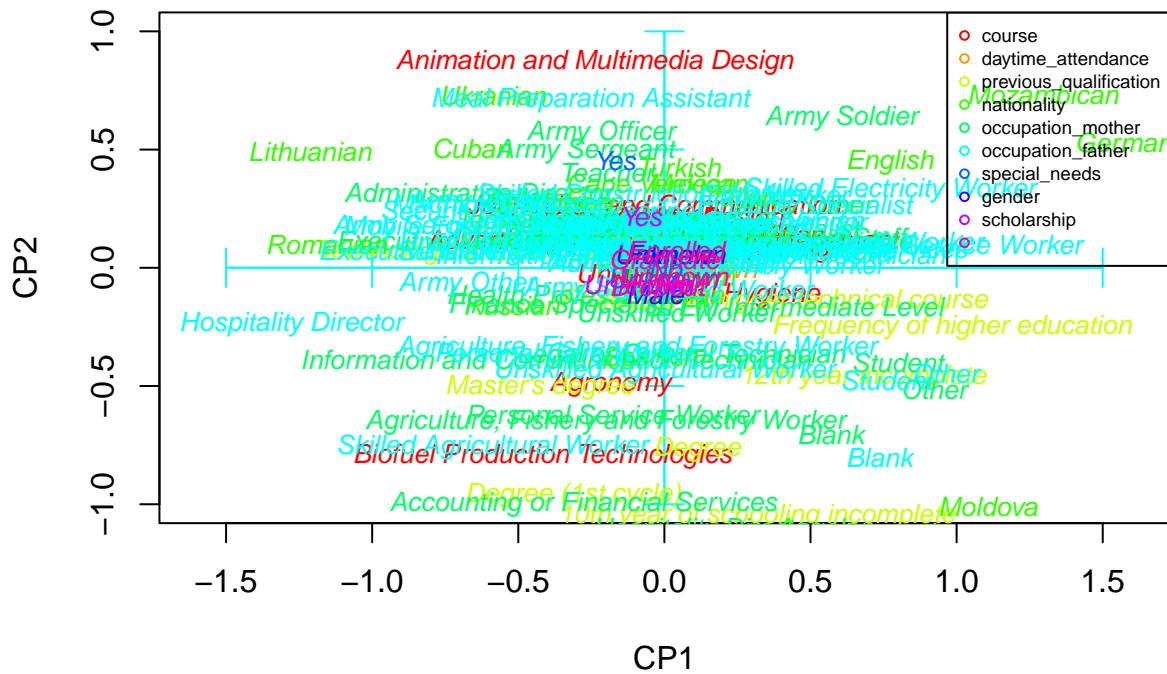
En el primer gràfic es veu que els individus es concentren en els valors més petits de PC1 i PC2. Passa el mateix quan es mira PC3 amb PC2, PC4 amb PC2 i PC5 amb PC2. Així mateix, els individus es distribueixen bastant uniformement en els darrers dos gràfics (PC5-PC3 i PC5-PC4).

## 5.2 Projecció de les variables categòriques

Variables categòriques

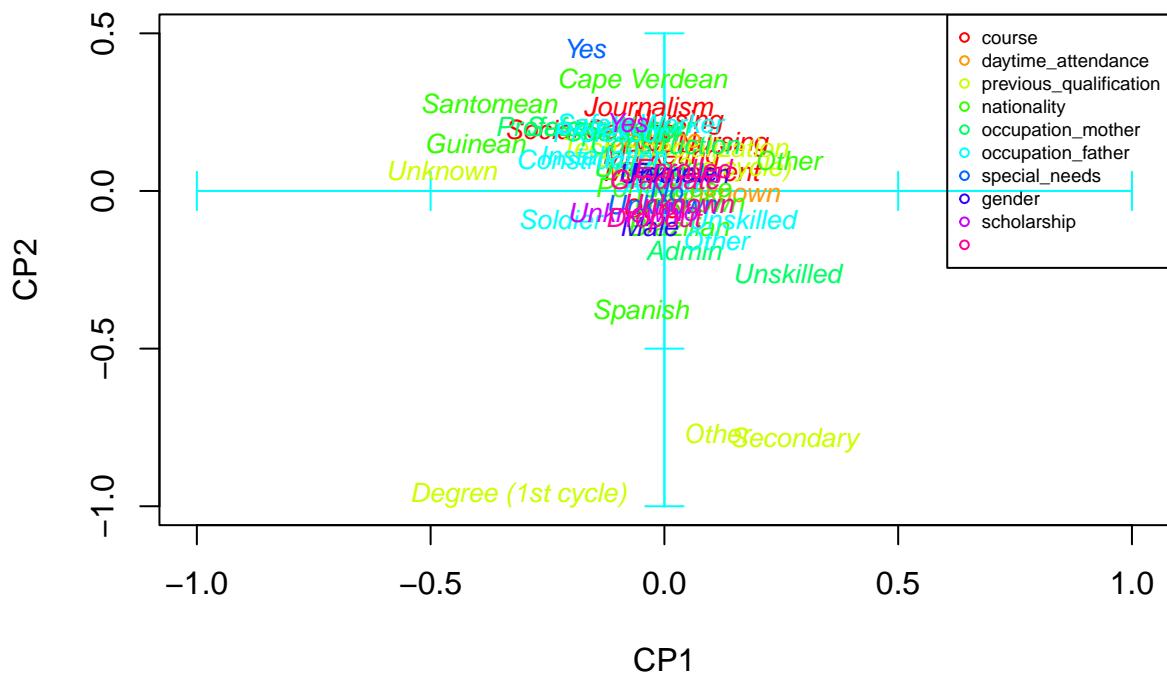
```
##          course      daytime_attendance previous_qualification
##            1                  2                      3
## nationality      occupation_mother      occupation_father
##            4                  5                      6
## special_needs      gender      scholarship
##            7                  8                      9
##          target
##            18
```

Si en el mateix gràfic s'afegeixen totes les modalitats de totes les variables categòriques, s'obté el següent:



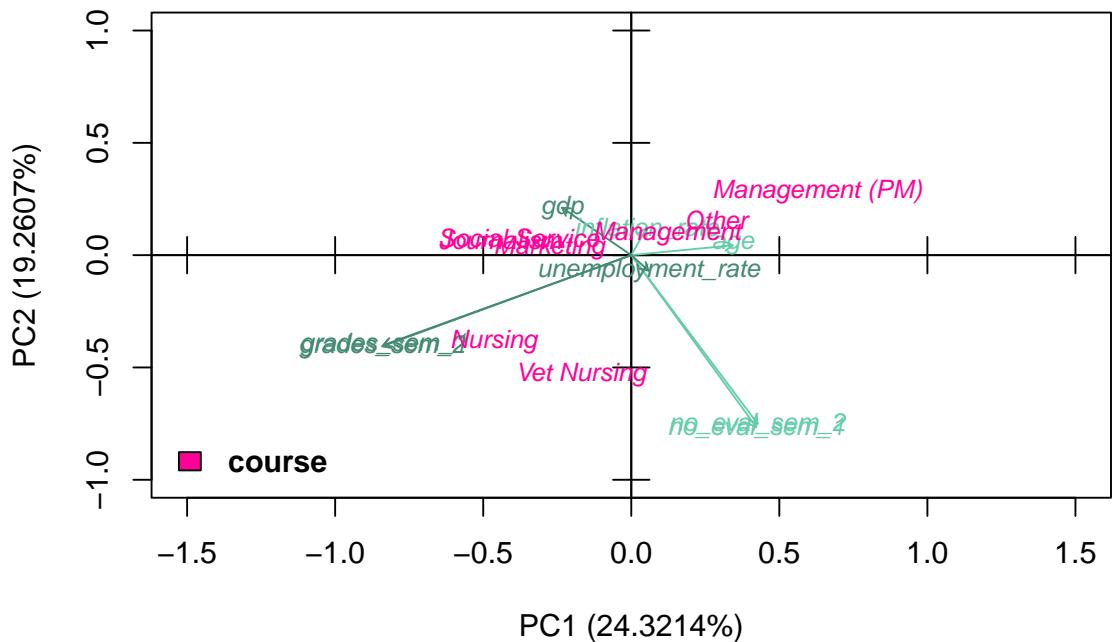
Aquest gràfic no es pot interpretar, ja que en tenir tantes modalitats aquestes no es poden distingir.

Es limita el nombre màxim de modalitats a 8 per variable. A més a més, s'escurcen els noms d'algunes modalitats per tal que siguin llegibles al gràfic. Es repeteix aquest procediment per a les següents variables categòriques: course, previous qualification, nationality, occupation mother i occupation father, que són les que tenen més modalitats.

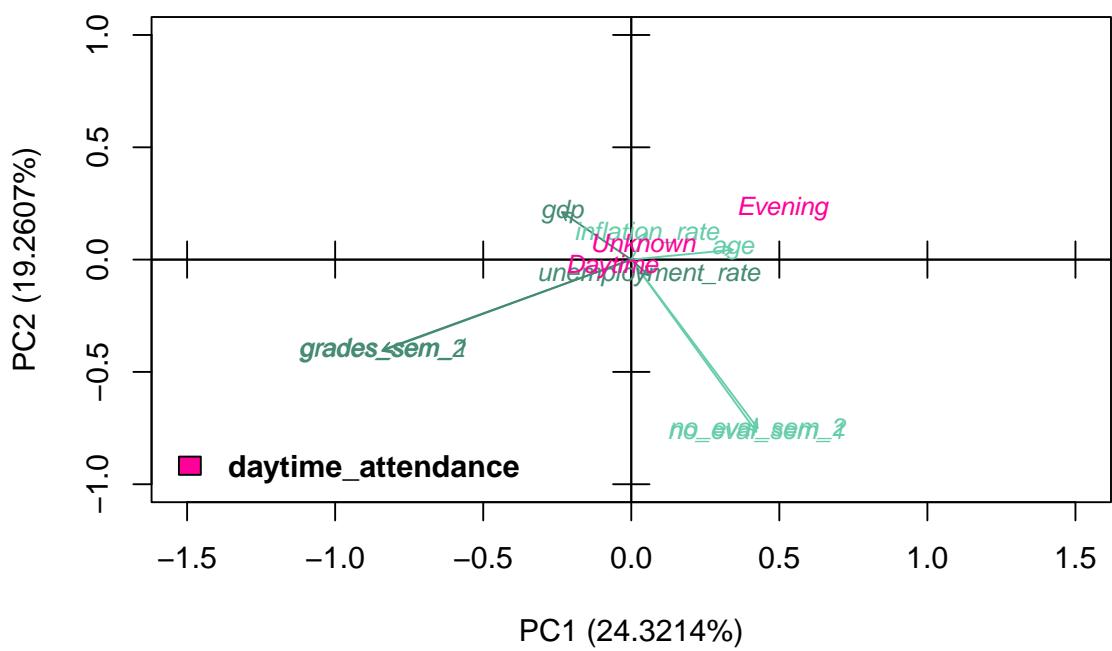


Es crea un gràfic per a cada variable categòrica:

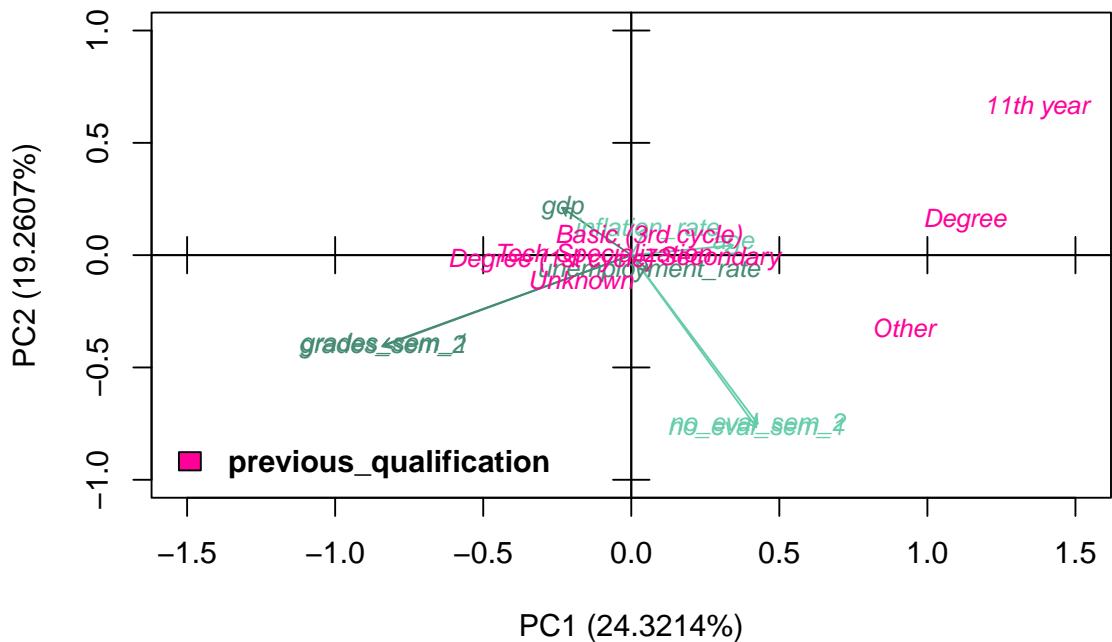
### Projeccions sobre el pla factorial de variables categòriques



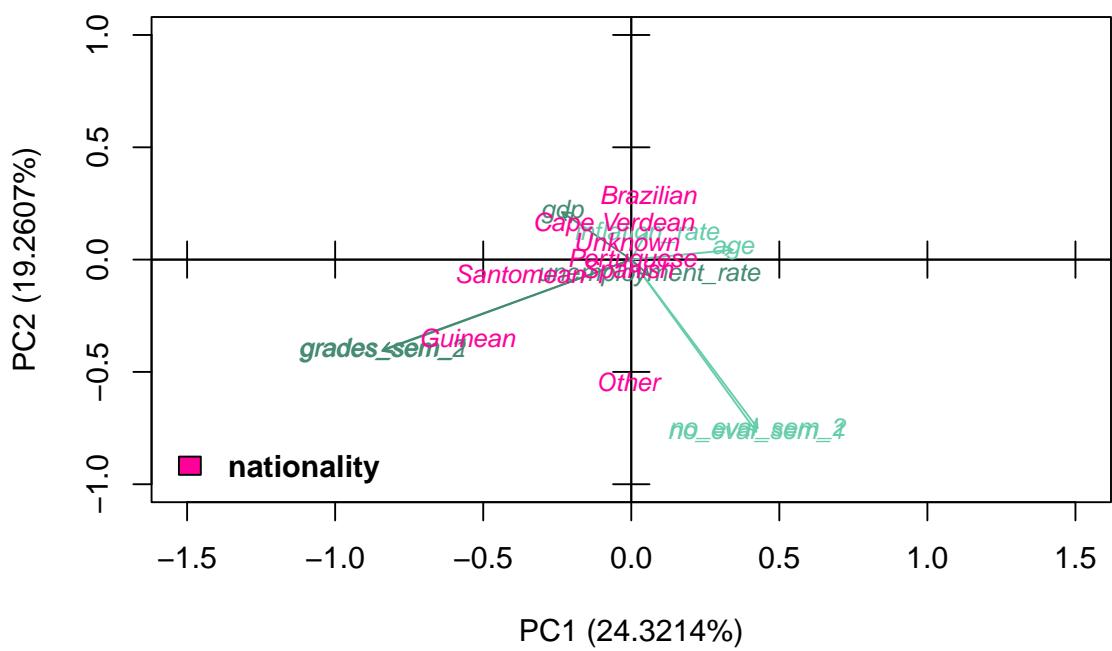
### Projeccions sobre el pla factorial de variables categòriques



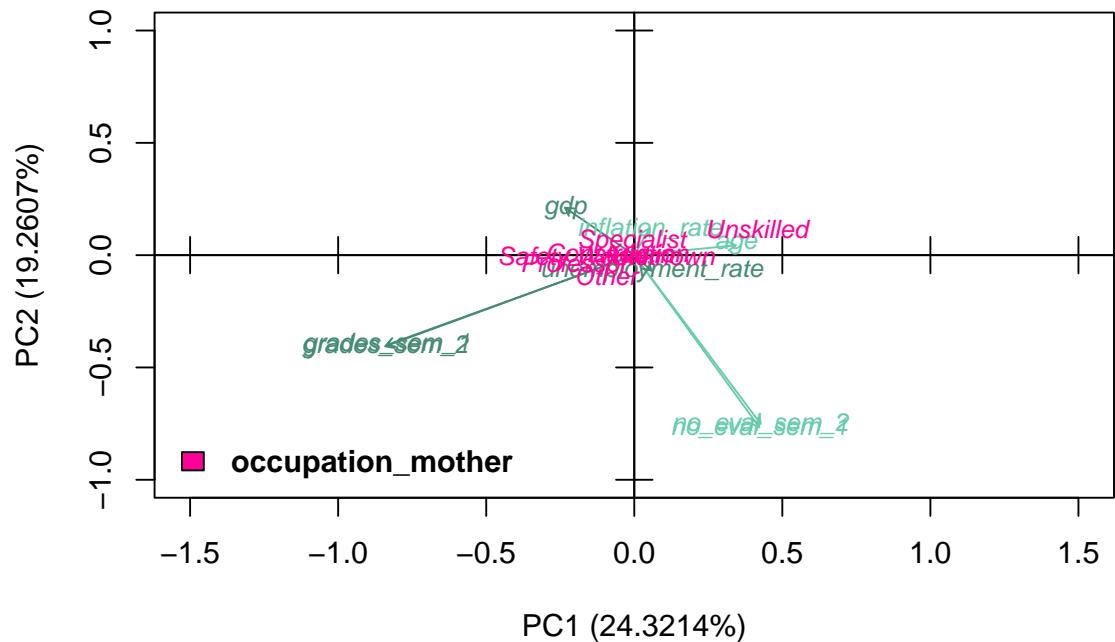
### Projeccions sobre el pla factorial de variables categòriques



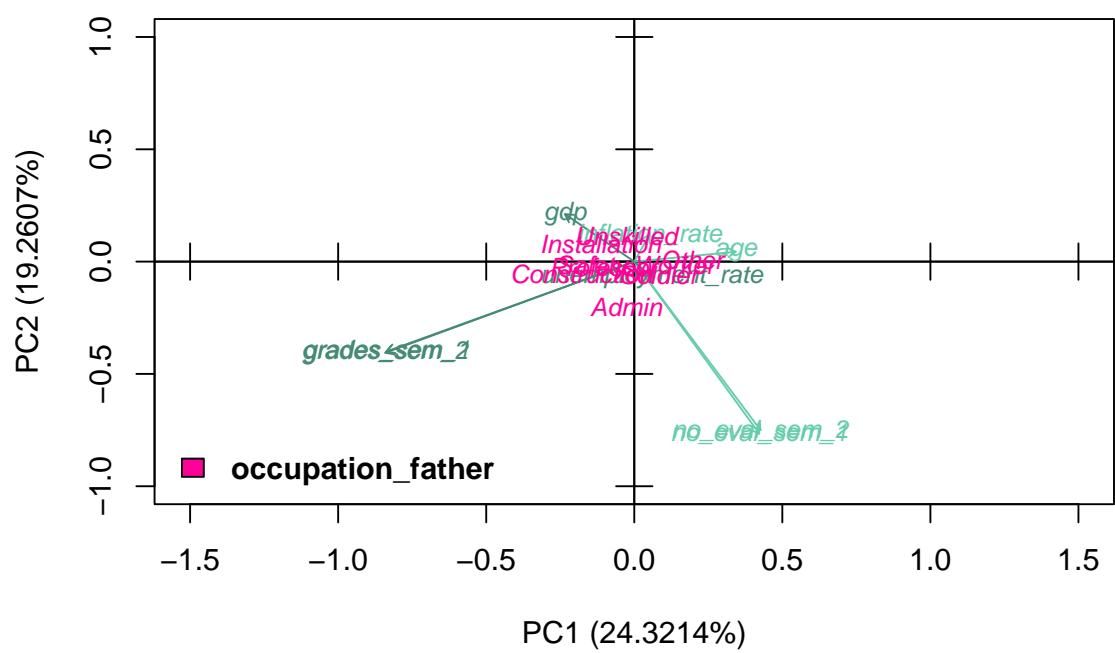
### Projeccions sobre el pla factorial de variables categòriques



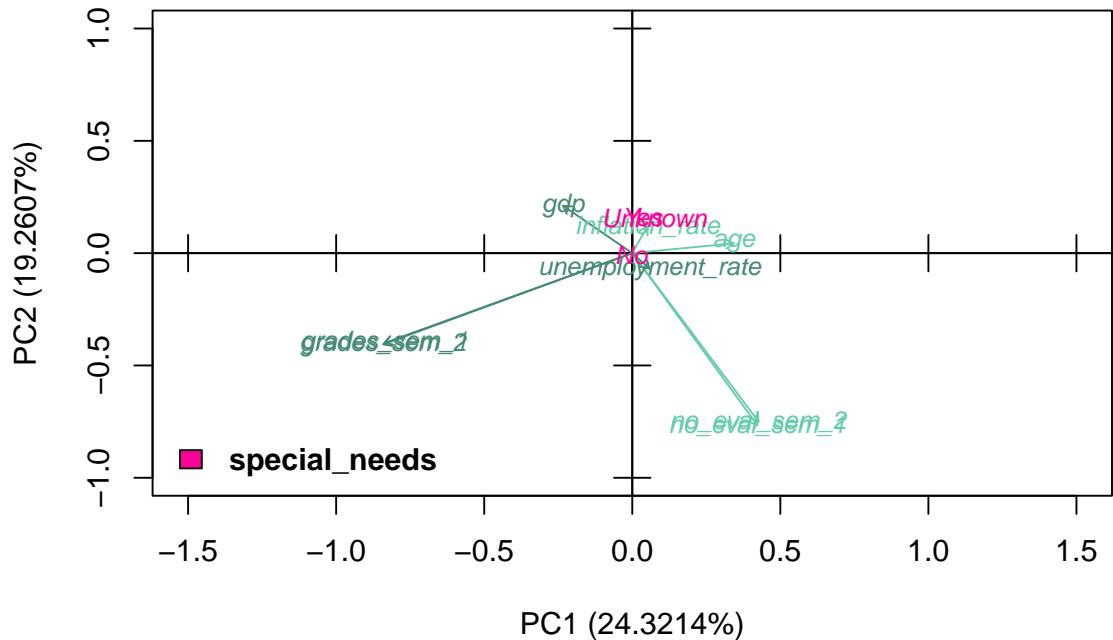
### Projeccions sobre el pla factorial de variables categòriques



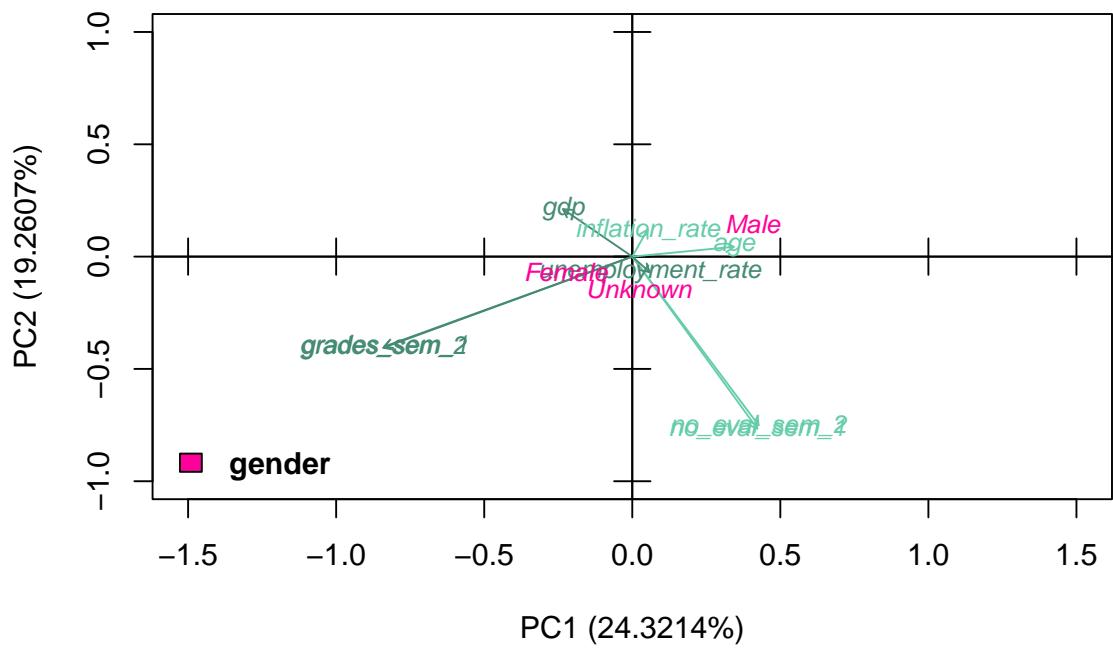
### Projeccions sobre el pla factorial de variables categòriques



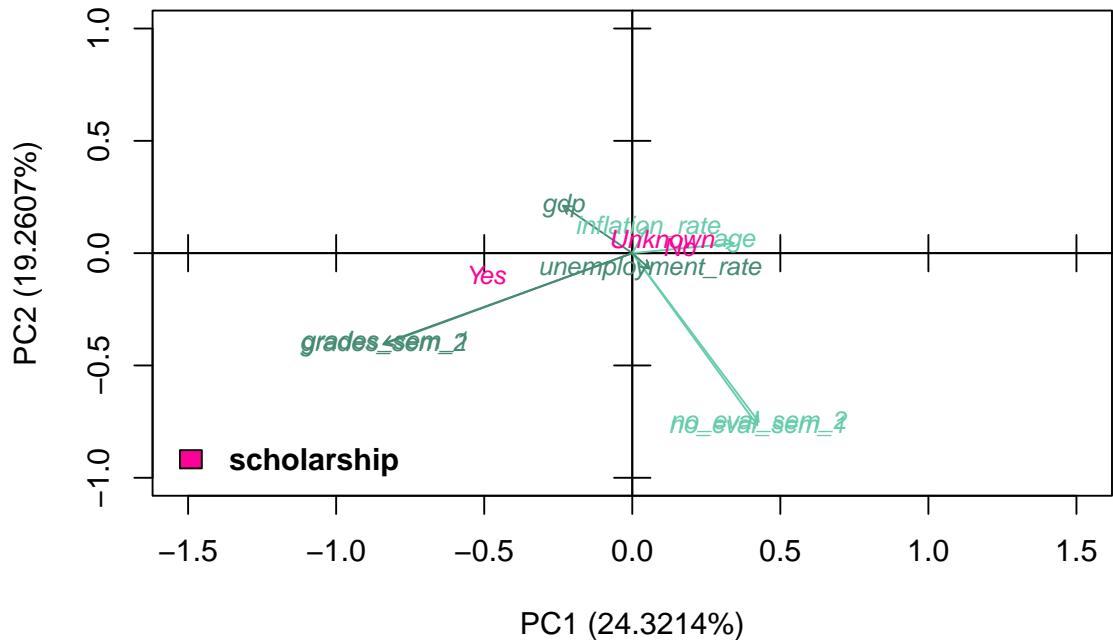
### Projeccions sobre el pla factorial de variables categòriques



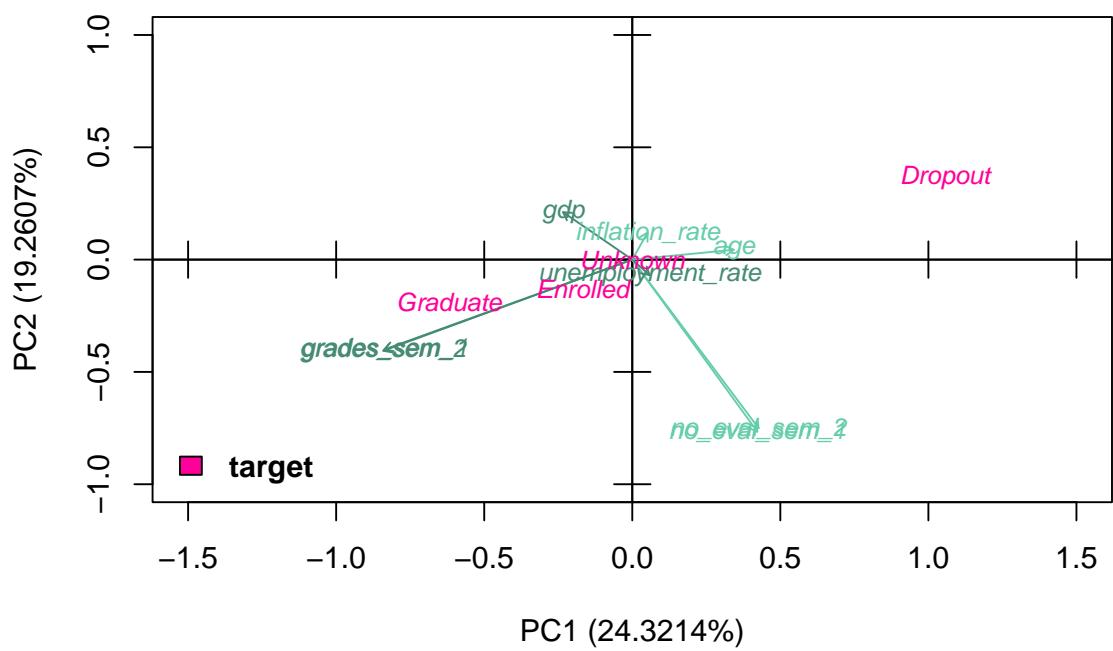
### Projeccions sobre el pla factorial de variables categòriques



### Projeccions sobre el pla factorial de variables categòriques



### Projeccions sobre el pla factorial de variables categòriques



Per a la variable **course**, es veu que la majoria de les modalitats de la variable es troben molt a prop de l'origen, fet que implica que expliquen les dues components de manera similar però amb una baixa contribució

sobre les components. Destaca la modalitat Vet Nursing, que explica la segona component, ja que es troba tot just a sobre de l'eix d'ordenades. Nursing té angles molt similars amb la primera i la segona component, per tant, es pot dir que les expliquen d'igual manera.

En la variable **daytime\_attendance**, es veu que la modalitat Evening sembla explicar totes dues components per igual.

La modalitat Degree de la variable **previous\_qualification** explica més la primera component que la segona. Per aquesta variable, la majoria de les modalitats estan molt propers a l'origen. 11th year sembla explicar totes dues components per un igual.

Per a la variable **nationality**, la modalitat Other explica perfectament la segona component, ja que es troba exactament a sobre de l'eix d'ordenades. La modalitat Guinean forma un angle similar amb totes dues components. La majoria de les modalitats de la variable **occupation\_mother** expliquen totes dues components per igual, però hi tenen una contribució molt baixa, ja que es troben molt properes a l'origen.

El mateix es pot dir per a la variable **occupation\_father** excepte per la modalitat Admin, que en trobar-se sobre l'eix d'ordenades i no sobre el de coordenades explica la segona component.

Pel que fa a les modalitats de la variable **special\_needs**, es veu que només la modalitat No es troba sobre l'origen i que Yes i Unknown només expliquen la segona component.

Per a la variable **gender**, Unknown explica la segona component. Female, en tenir un angle similar entre la primera i la segona es pot dir que les explica per igual.

La modalitat No de la variable **scholarship** explica la primera component a trobar-se just a sobre de l'eix de coordenades. No obstant això, té una contribució molt baixa sobre aquesta component, ja que es troba bastant propera a l'origen.

Per a la variable **target** destaca la modalitat Dropout, la qual té una contribució força elevada sobre totes dues components. Enrolled sembla explicar més la segona component que la primera.

## 6 ACM

Per començar s'han d'escollir quines variables seran les variables actives, i quines les complementàries:

### Categòriques actives:

- scolarship
- target
- gender
- daytime\_attendance
- occupation\_mother
- course

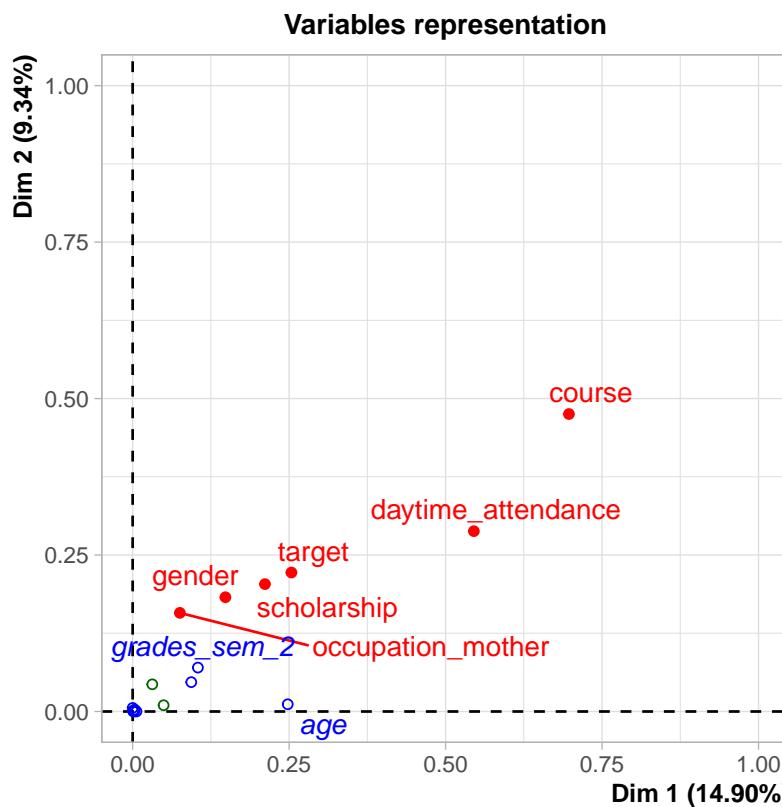
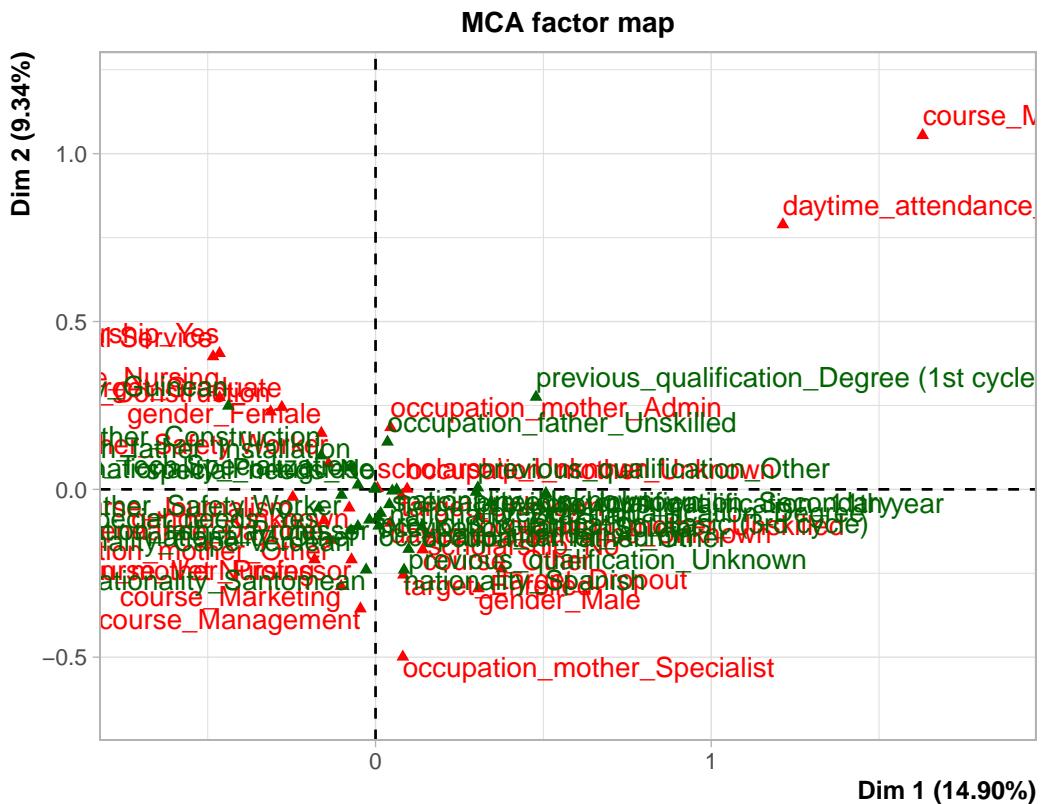
### Categòriques suplementàries:

- Nationality
- Special\_needs
- occupation\_father
- previous\_qualification

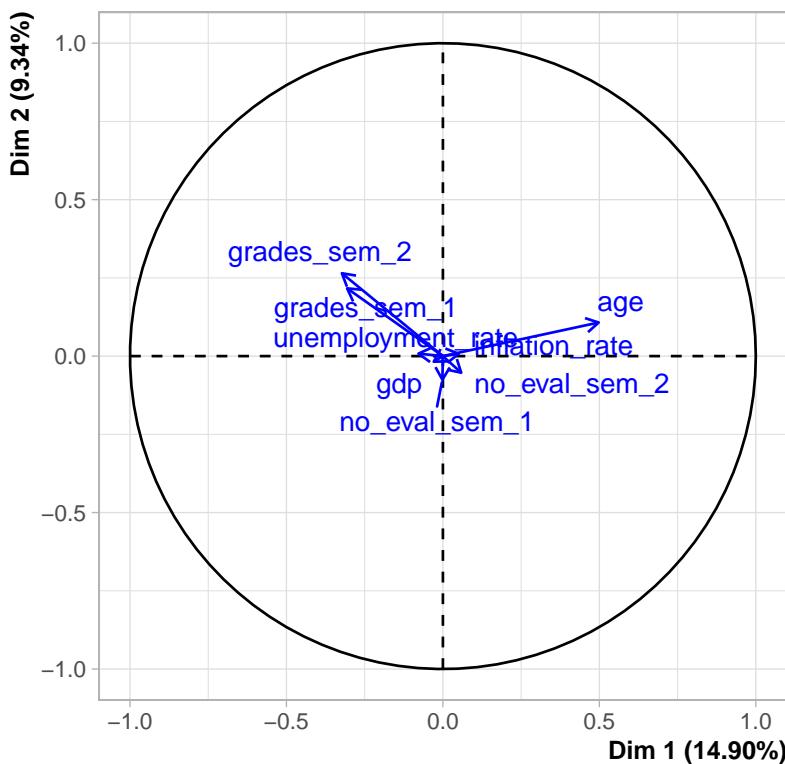
### Numèriques suplementàries:

Com en l'ACP es va decidir no eliminar cap variable numèrica, s'utilitzaran totes les variables numèriques com a variables suplementàries.

A partir d'aquesta selecció es procedirà a realitzar l'ACM segons el mètode de Burt. Es comença quantificant la inèrcia de cada dimensió a partir de la variància explicada de cada una.



### Supplementary quantitative variables



```

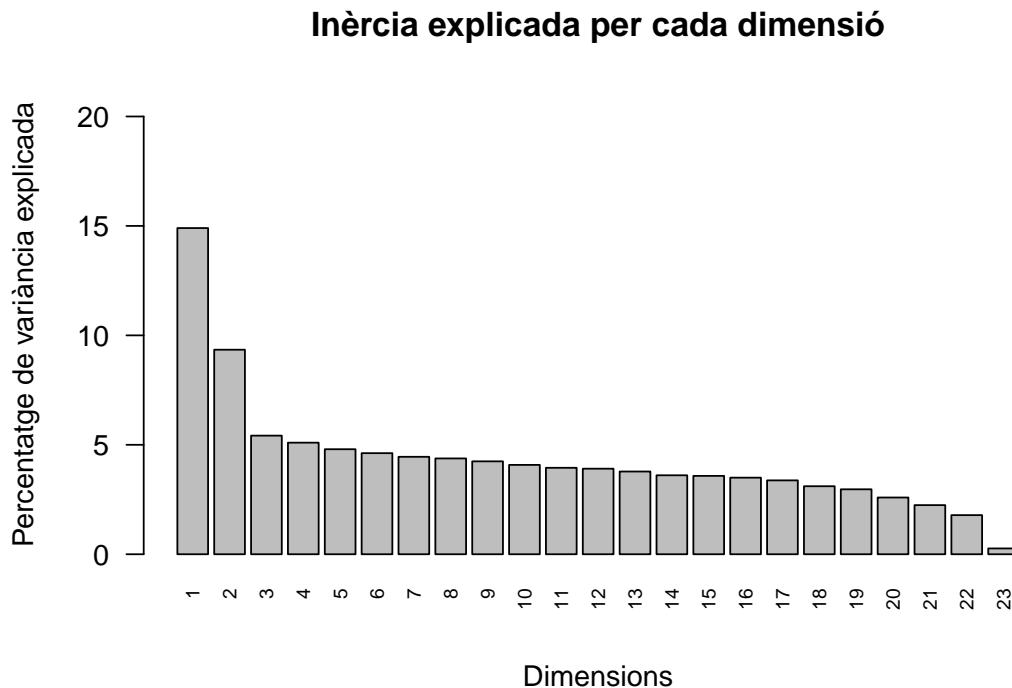
## **Results of the Multiple Correspondence Analysis (MCA)**
## The analysis was performed on 4424 individuals, described by 18 variables
## *The results are available in the following objects:
##
##      name           description
## 1  "$eig"          "eigenvalues"
## 2  "$var"          "results for the variables"
## 3  "$var$coord"    "coord. of the categories"
## 4  "$var$cos2"     "cos2 for the categories"
## 5  "$var$contrib"  "contributions of the categories"
## 6  "$var$v.test"   "v-test for the categories"
## 7  "$ind"          "results for the individuals"
## 8  "$ind$coord"    "coord. for the individuals"
## 9  "$ind$cos2"     "cos2 for the individuals"
## 10 "$ind$contrib"  "contributions of the individuals"
## 11 "$quanti.sup"   "results for the supplementary quantitative variables"
## 12 "$quanti.sup$coord" "coord. of the supplementary quantitative variables"
## 13 "$quali.sup"    "results for the supplementary categorical variables"
## 14 "$quali.sup$coord" "coord. for the supplementary categories"
## 15 "$quali.sup$cos2" "cos2 for the supplementary categories"
## 16 "$quali.sup$v.test" "v-test for the supplementary categories"
## 17 "$call"          "intermediate results"
## 18 "$call$marge.col" "weights of columns"
## 19 "$call$marge.li"  "weights of rows"

```

En els gràfics obtinguts es pot veure la variabilitat que expressen cada una de les variables categòriques en funció de les dimensions 1 i 2. Aquelles variables que estiguin més a prop de l'origen de coordenades

aporten molt poca informació respecte a la variabilitat de les dades i, per tant, són poc importants. En canvi, aquelles variables més allunyades del centre aporten informació més rellevant, com l'horari de classes o el curs.

Es representen gràficament la inèrcia que explica cada una de les dimensions generades:



Si una dimensió té una inèrcia baixa, significa que totes les modalitats estan molt properes al centre de gravetat, i en conseqüència, són molt similars. A mesura que augmenta la inèrcia, va augmentant la distància al centre de gravetat i, per tant, es redueix la similitud.

Per poder estudiar-ho més a fons, es realitza la següent taula on es pot observar per a cada dimensió, el seu valor propi, el percentatge de variància (o inèrcia) explicada, i el percentatge de variància (o inèrcia) acumulada:

##	eigenvalue	percentage of variance	cumulative percentage of variance
## dim 1	0.10	14.90	14.90
## dim 2	0.06	9.34	24.25
## dim 3	0.04	5.42	29.67
## dim 4	0.04	5.10	34.76
## dim 5	0.03	4.80	39.56
## dim 6	0.03	4.62	44.18
## dim 7	0.03	4.45	48.63
## dim 8	0.03	4.38	53.01
## dim 9	0.03	4.24	57.25
## dim 10	0.03	4.08	61.34
## dim 11	0.03	3.95	65.28
## dim 12	0.03	3.91	69.19
## dim 13	0.03	3.78	72.97
## dim 14	0.03	3.61	76.58
## dim 15	0.02	3.58	80.16

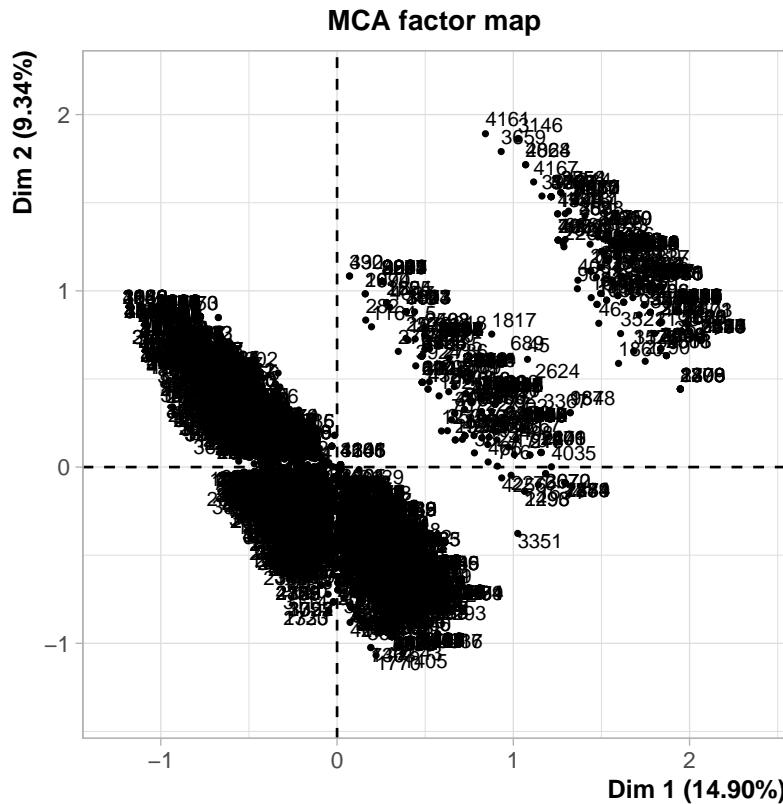
## dim 16	0.02	3.50	83.66
## dim 17	0.02	3.38	87.04
## dim 18	0.02	3.11	90.14
## dim 19	0.02	2.97	93.11
## dim 20	0.02	2.59	95.70
## dim 21	0.02	2.25	97.95
## dim 22	0.01	1.79	99.73
## dim 23	0.00	0.27	100.00

Tenim un total de 23 dimensions. Es veu que la dimensió 1 és la que destaca més per sobre la resta, explicant un 14.90% de la variabilitat de les dades, seguida de les dimensions 2, que també destaca per sobre de la resta, explicant un 9.34% de la variabilitat de les dades. A partir de la dimensió 3, es veu que el gràfic s'estabilitza bastant ja fins a l'última dimensió.

Per tant, en total les dues primeres dimensions ja expliquen un 24.25% de la variabilitat de les dades, i es necessiten 15 dimensions per arribar a tenir una inèrcia acumulada per sobre del 80%.

## 6.1 Plot individus

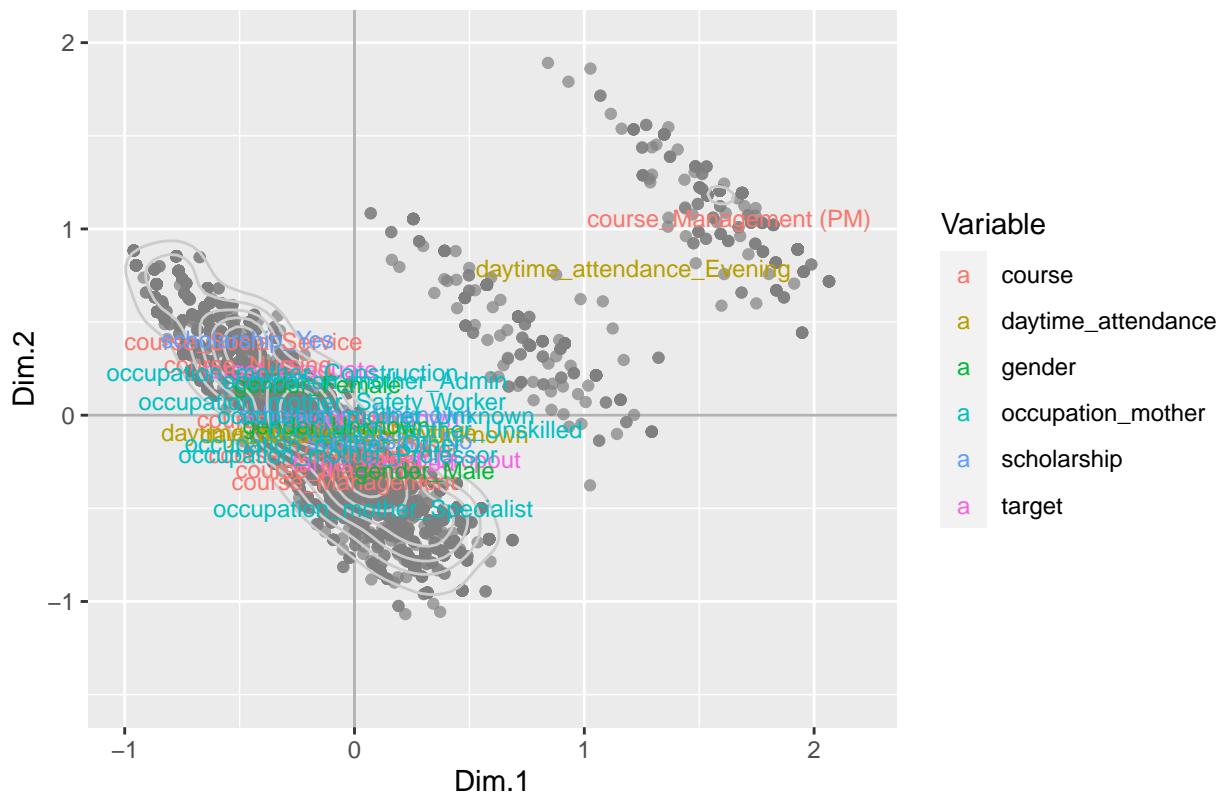
Es representa gràficament com es distribueixen els individus en funció de les dues primeres dimensions que expliquen un 24.25% de la variabilitat:



A simple vista, es pot veure que els individus es diferencien en 3 grups. Tot i això, la gran majoria d'individus es concentren molt propers a l'origen de manera que aporten poca informació sobre quines variables rellevants tenen respecte als altres individus. Es passa a estudiar cada variable per observar si hi ha algun tipus d'associació entre elles.

## 6.2 Plot variables

MCA plot of variables using R package FactoMineR

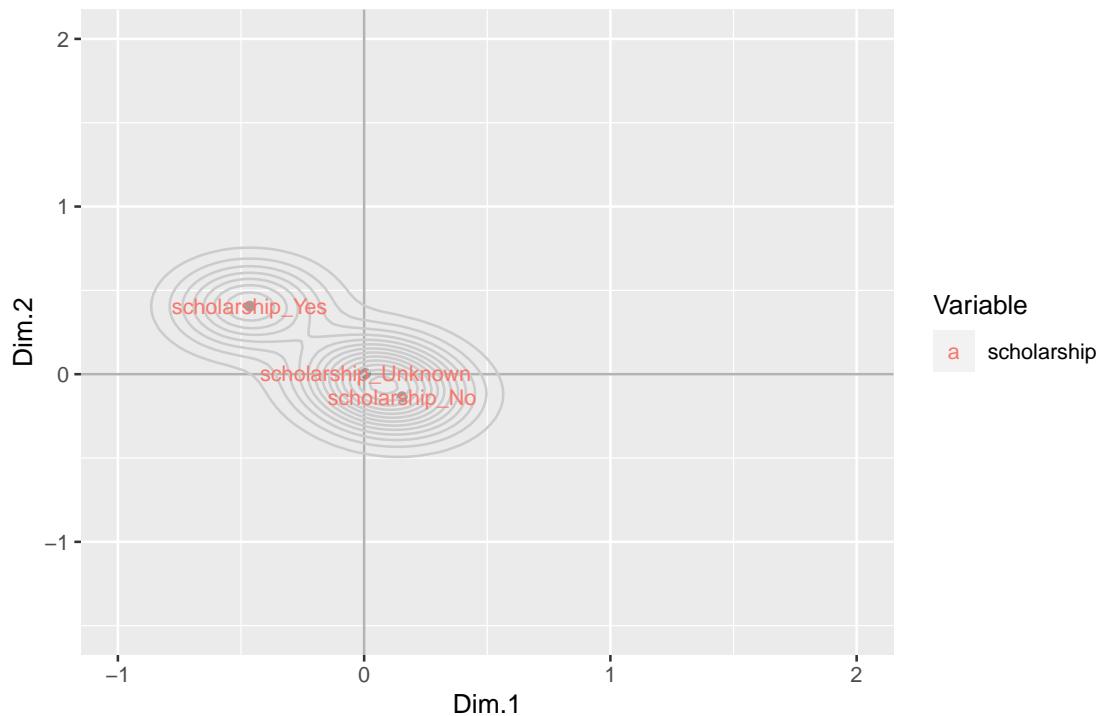


Igual que en el gràfic per individus, es veu que les dades es divideixen en 3 grups. Es poden classificar segons els eixos com:

- Grup 1:  $X = [-1, 0.5]$ ,  $Y = [-1, 1]$
- Grup 2:  $X = [0, 1]$ ,  $Y = [-0.5, 1]$
- Grup 3:  $X = [1, 2]$ ,  $Y = [0.5, 2]$

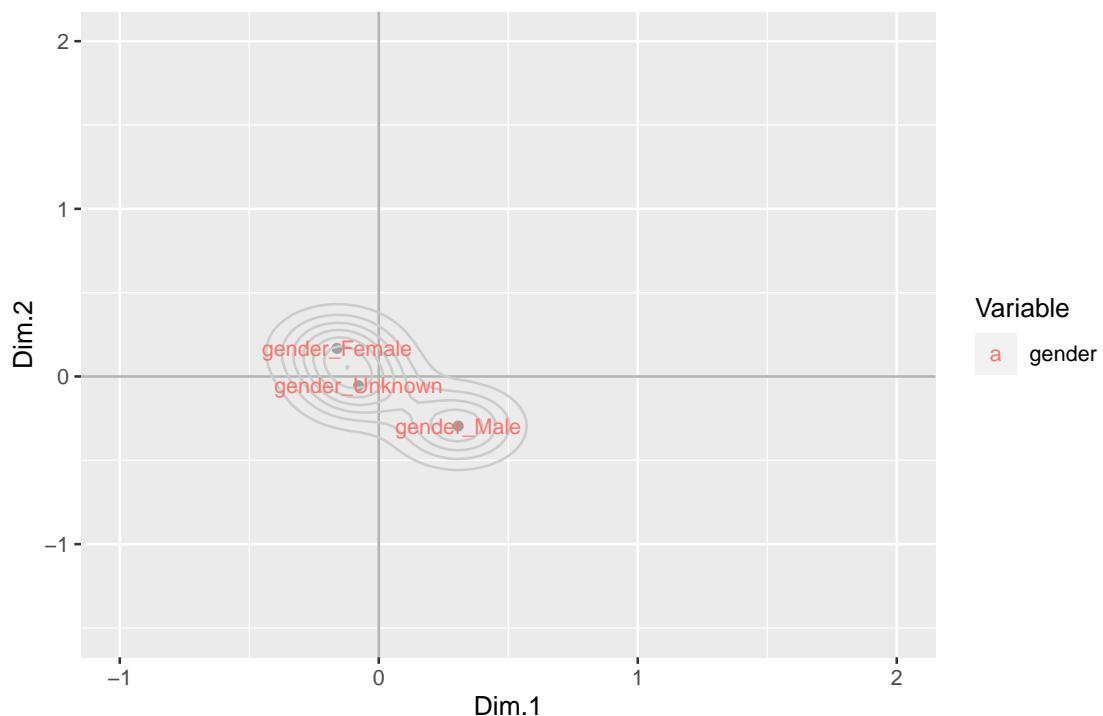
Amb totes les variables actives juntes, és impossible diferenciar-les i poder extreure'n conclusions, per això s'aniran fent gràfics de les diferents variables:

MCA plot of variables using R package FactoMineR



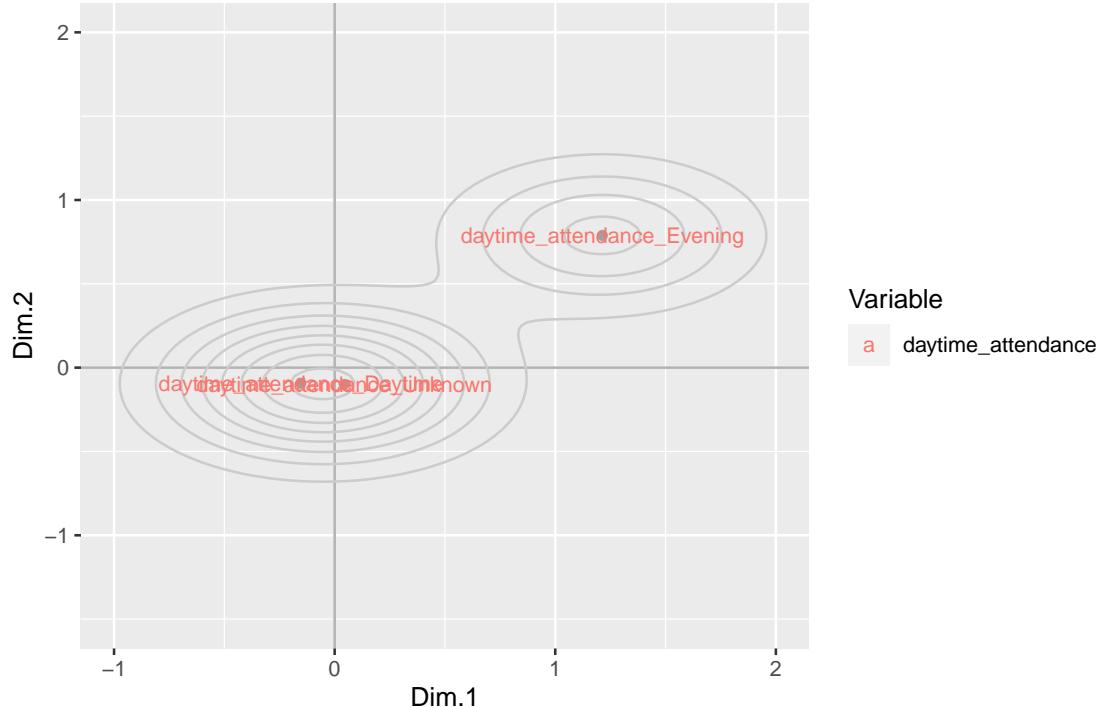
A simple vista, es pot observar que totes les modalitats de la variable *Scholarship* no estan molt repartides en el gràfic. Es pot veure que les modalitats ‘Unknown’ i ‘No’ no expliquen gaire variabilitat en situar-se molt a prop del centre de gravetat de les dimensions. De totes maneres, la categoria ‘Yes’ sí que sembla que podria tenir certa rellevància explicativa (II quadrant).

MCA plot of variables using R package FactoMineR



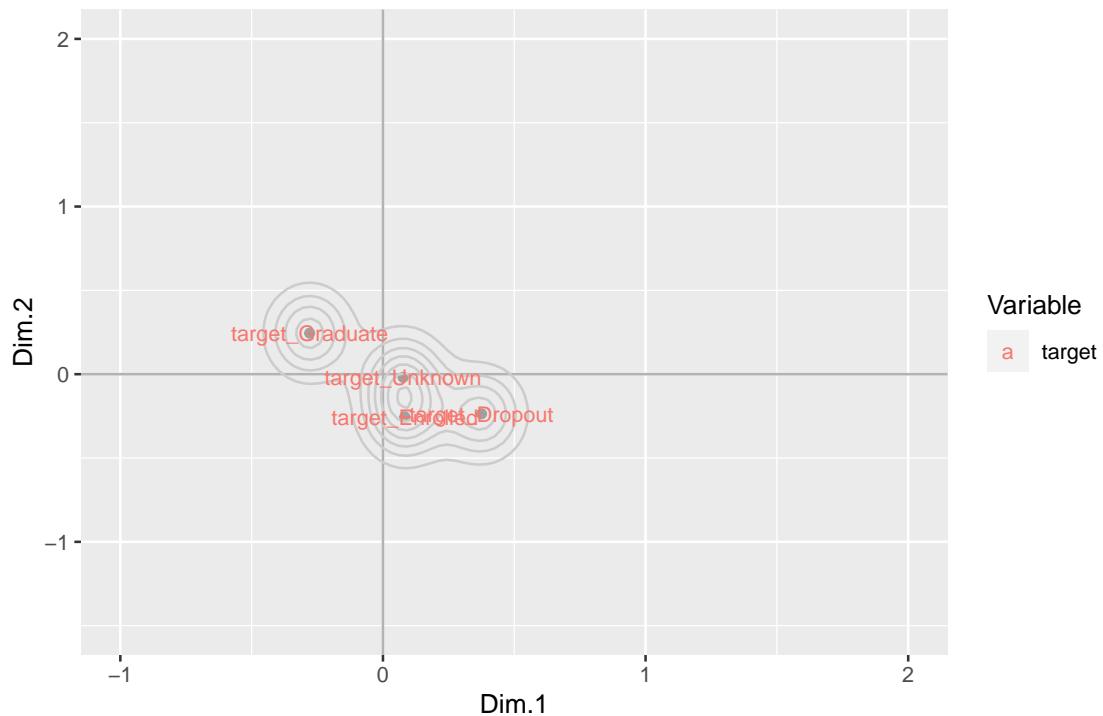
Com en la variable *Scholarship*, en la variable *Gender* s'observa que totes les seves modalitats queden agrupades al voltant de l'origen de coordenades. En aquest cas les modalitats 'Female' i 'Unknown' no aporten informació que pugui ser útil, mentre que la categoria 'Male' queda una mica més lluny (quadrant IV) i sembla que podria tenir certa associació.

MCA plot of variables using R package FactoMineR



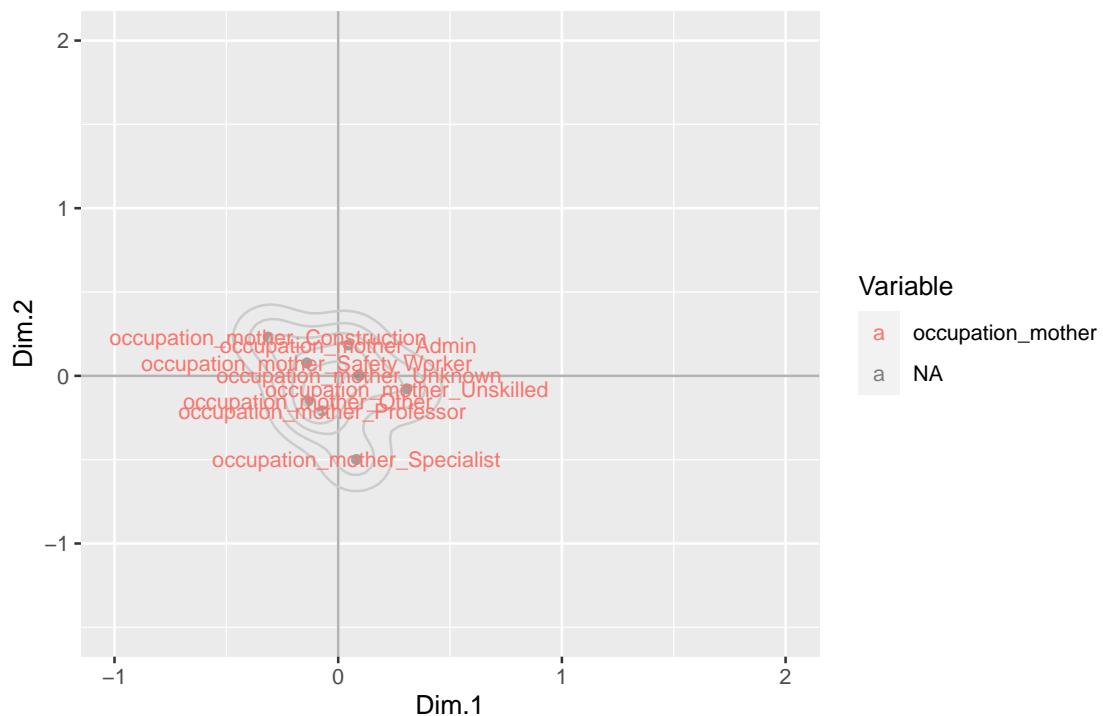
Pel que fa a la variable *daytime\_attendance* es veu que les modalitats es reparteixen d'una forma força interessant en el gràfic. Si bé les modalitats 'Daytime' i 'Unknown' expliquen poca variabilitat en quedar a prop del centre, la categoria 'Evening' queda més separada i se situa en el quadrant I de forma clara.

MCA plot of variables using R package FactoMineR



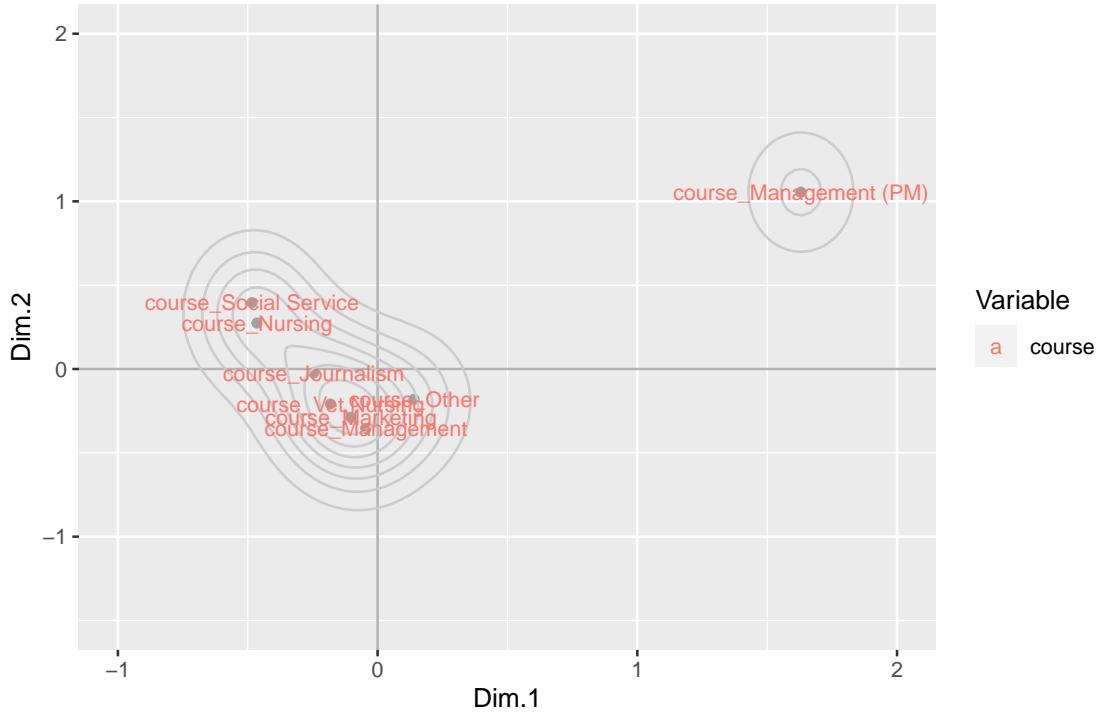
Altra vegada, es veu que totes les modalitats de la variable *Target* es concentren al voltant del centre i sense gaire variabilitat. Tot i això, hi ha dues modalitats que sembla que podrien aportar certa informació. La categoria ‘Dropout’ se situa de forma més clara al quadrant IV i la categoria ‘Graduate’ se situa al quadrant II.

MCA plot of variables using R package FactoMineR



Observant la variable *Occupation\_mother* es veu que totes les modalitats d'aquesta variable s'agrupen al voltant del centre. Si bé hi ha algunes variables que es poden situar de forma més clara en un dels quadrants o lluny de l'origen, en quedar totes molt a prop entre elles no hi ha prou certesa per assegurar que aporten informació verídica.

MCA plot of variables using R package FactoMineR



Pel que fa a la variable *Course*, es veu que gran part de les seves modalitats s'agrupen al centre menys una. La categoria ‘Management(PM)’ se situa lluny de totes les altres i en el quadrant I. D'altra banda, les modalitats ‘Social Service’ i ‘Nursing’ si situen també lleugerament lluny del centre i en el quadrant II.

### 6.3 Associacions

Havent estudiat cada una de les variables per separat s'ha pogut veure que hi ha certes modalitats de diferents variables que se situen en una mateixa zona del gràfic. Això significa que pot haver-hi certa associació entre aquestes modalitats, tot i que no es pot entrar a definir quin tipus de relació es dona entre elles. També s'ha de tenir en compte que havent utilitzat aquestes dues dimensions només es pot explicar un percentatge reduït de la variabilitat (24.25%). Així doncs, si es tinguessin les eines per poder utilitzar més dimensions, es podrien obtenir uns resultats més verídics i lleugerament diferents.

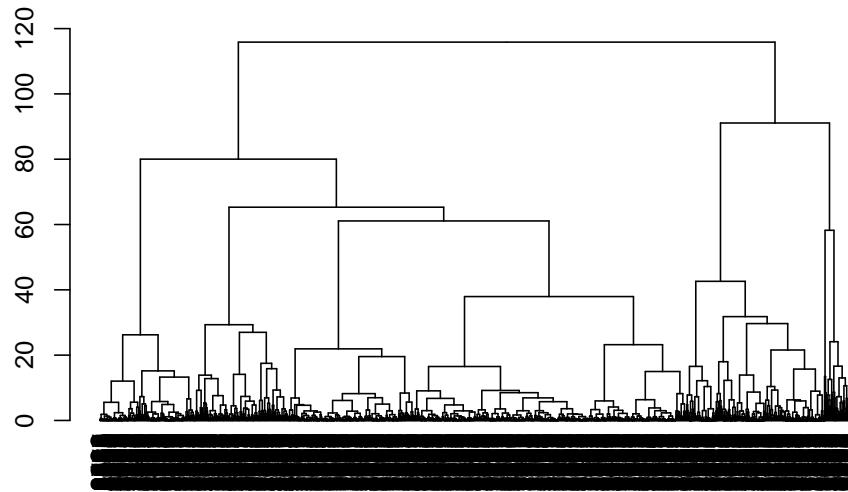
Les associacions que s'han trobat són les següents:

- En el quadrant I hi coincideixen aquells individus que van de tardes i els que estudien el curs de Management.
- En el quadrant II es troben aquells estudiants que estudien Social Service o Nursing, que han aconseguit graduar-se i han tingut una beca.
- En el quadrant IV es troben generalment els estudiants homes i els que han abandonat la carrera.

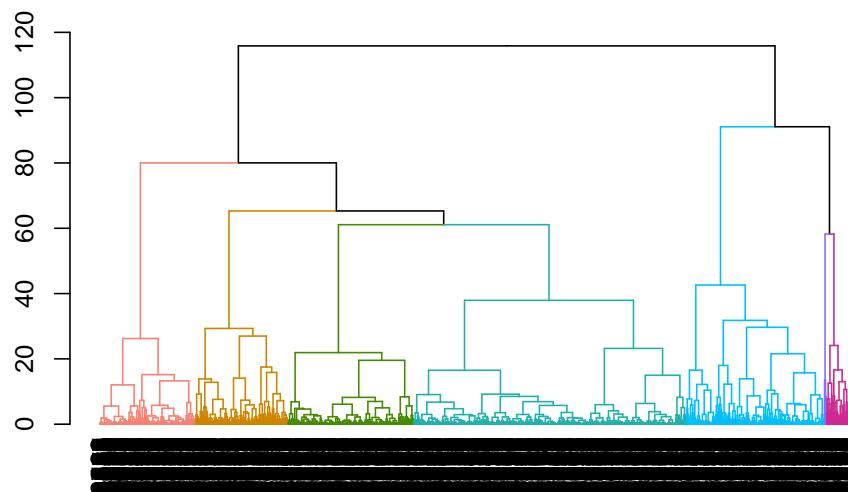
## 7 Clustering jeràrquic sobre les components factorials retingudes a l'ACP

Es fa el clustering per l'ACP. Tenim les coordenades de cada individu en les 5 dimensions que comprenen un 80% de la inèrcia. Amb aquestes dades es calcula la distància euclidiana, ja que són dades numèriques, s'aplica el mètode de Ward per fer el clustering, i posteriorment graficar el dendrograma.

Es veu clarament que el nombre òptim de grups per classificar les dades és 7:



Es visualitzen els 7 clústers:



## 7.1 Profiling post ACP i ACM

Com ja s'ha mencionat a l'apartat de profiling, s'elaboren els perfils per poder diferenciar entre grups amb la distinció adequada. S'utilitzarà el mateix procés d'anàlisi i les mateixes eines gràfiques.

Primerament, s'estudia el p-valor de les variables numèriques, observant quines són significatives, amb el nivell de significació establert en  $\alpha = P(\text{error tipus I}) = 0.05$ .

```
##          age grades_sem_1 no_eval_sem_1 grades_sem_2 no_eval_sem_2
## 1 5.986680e-17 0.000000e+00 3.004941e-07 0.000000e+00 8.541882e-08
## 2 0.000000e+00 1.467645e-39 4.121801e-01 1.929312e-48 3.662833e-04
## 3 0.000000e+00 1.472747e-17 4.184462e-03 9.429821e-18 4.543281e-03
## 4 1.663690e-04 1.864848e-05 3.501934e-160 9.995098e-06 4.421445e-273
## 5 2.374741e-08 2.312979e-28 2.571230e-01 1.144361e-31 1.217628e-04
## 6 0.000000e+00 1.181729e-99 0.000000e+00 3.100684e-122 1.673839e-11
## 7 5.717593e-02 8.585743e-08 0.000000e+00 5.045473e-05 3.786235e-298
##      unemployment_rate inflation_rate           gdp
## 1      2.632936e-01 1.210831e-03 3.171014e-02
## 2      1.583131e-53 0.000000e+00 0.000000e+00
## 3      4.182211e-06 2.109424e-15 3.222308e-01
## 4      3.067370e-01 2.540378e-04 1.612567e-08
## 5      2.915505e-111 2.016589e-280 0.000000e+00
## 6      0.000000e+00 1.751035e-02 1.668798e-275
## 7      3.417407e-01 1.066368e-01 3.112132e-03
```

Per mitjà del test chi-quadrat, de cadascuna respecte als clústers, es mira quines variables categòriques són significatives per definir els clústers.

```
##          course daytime_attendance previous_qualification
## 1 1.598026e-140 2.075290e-147 3.217078e-84
## 2 nationality occupation_mother occupation_father
## 3 7.947398e-01 3.304552e-26 1.724216e-12
## 4 special_needs gender scholarship
## 5 2.819927e-01 4.495225e-38 3.925393e-39
## 6 target
## 7 9.046327e-256
```

A la següent taula es mostra el valor 1 de les variables considerades com a significatives. Aquestes seran les estudiades gràficament.

```

##          course      daytime_attendance previous_qualification
##                1                      1                           1
##      nationality      occupation_mother      occupation_father
##                0                      1                           1
##      special_needs      gender      scholarship
##                0                      1                           1
##          age      grades_sem_1      no_eval_sem_1
##                1                      1                           0
##      grades_sem_2      no_eval_sem_2      unemployment_rate
##                1                      1                           0
##      inflation_rate      gdp      target
##                1                      0                           1

```

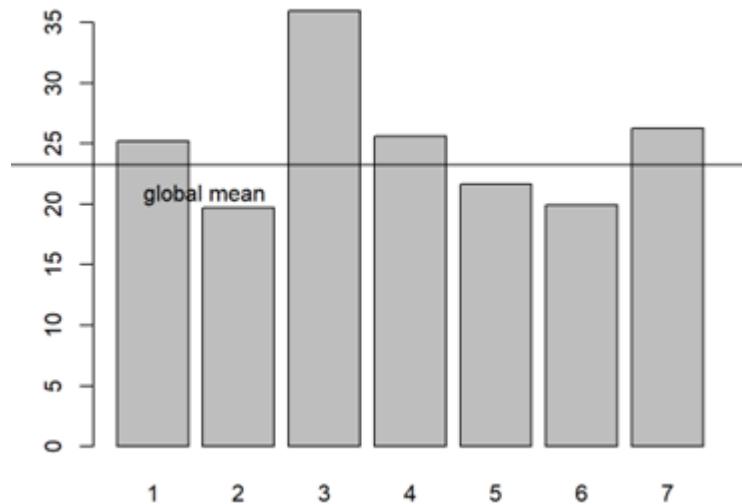
Es descarten les variables:

- Nacionalitat de l'estudiant
- Necessitats especials
- Unitats curriculars suspeses
- Taxa d'atur
- PIB

### 7.1.1 Anàlisi gràfica de les variables numèriques

*EDAT*

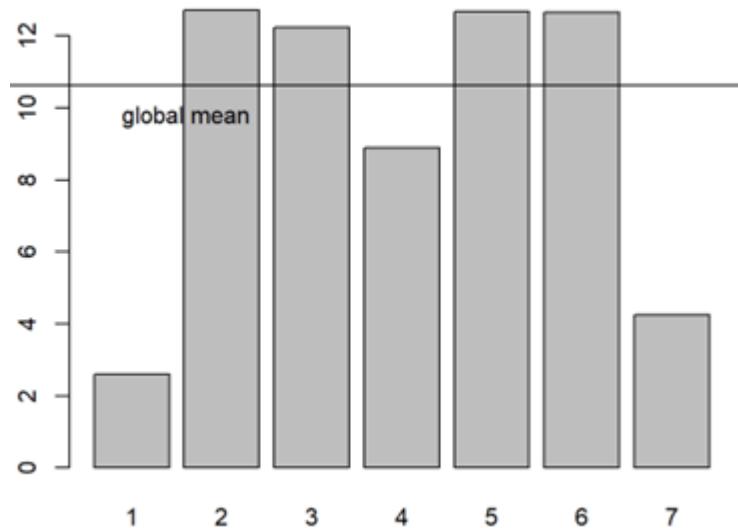
**Means of age by classe**



A aquest gràfic es veu com el clúster més allunyat de la mitjana és el tres, els grups dos i sis són els més joves, amb 20 anys de mitjana. La resta es troben al voltant de la mitjana. Es considera que són prou significatives les diferències per a explicar els grups.

*NOTES AL PRIMER SEMESTRE*

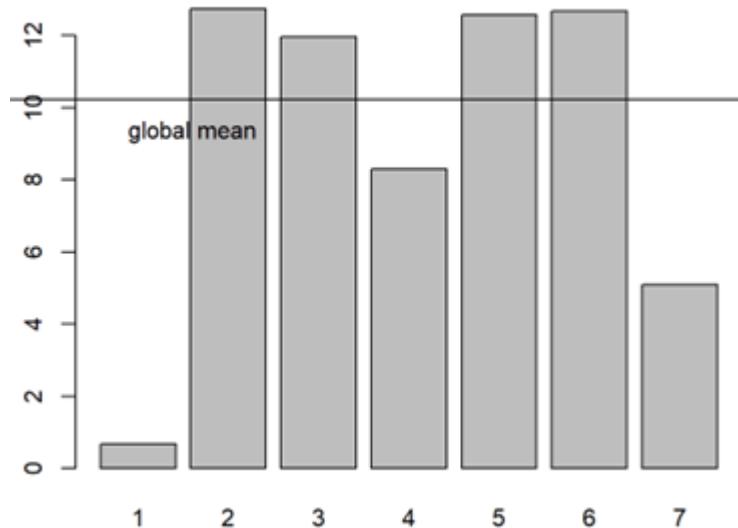
**Means of grades\_sem\_1 by classe**



Per a la variable que tracta sobre les notes al primer semestre, els clústers més significatius són l'1 i el 7 amb les notes mitjanes més baixes, suspeses. Les millors notes són dels grups 2, 3, 5 i 6, molt properes com per poder diferenciar. El grup 4 es troba per sota la mitjana, però propera a aquesta.

*NOTES AL SEGON SEMESTRE*

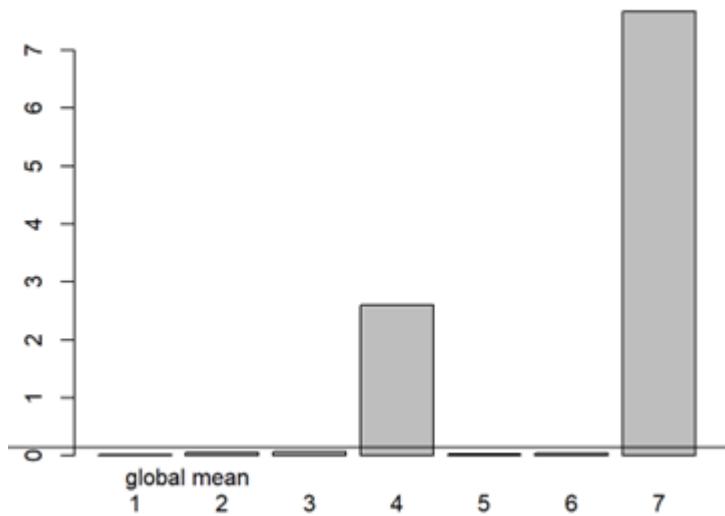
**Means of grades\_sem\_2 by classe**



S'observa com les notes al segon semestre segueixen la mateixa tendència que les notes al primer semestre. Parlant dels grups amb les notes més baixes, el grup 1 baixa la seva mitjana i el grup 7 la puja una mica, però no suficient per a apropar-se a la resta de clústers.

## UNITATS CURRICULARS SUSPESES

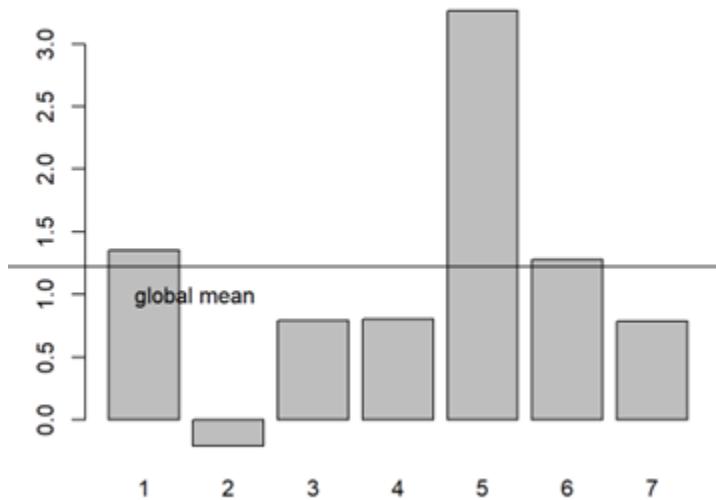
**Means of no\_eval\_sem\_2 by classe**



Parlant del nombre d'unitats curriculars suspeses al segon semestre, el clúster 7 és el que més unitats curriculars suspeses té, seguit pel grup 4. La resta de grups no tenen quasi unitats curriculars suspeses.

## TAXA D'INFLACIÓ

**Means of inflation\_rate by classe**



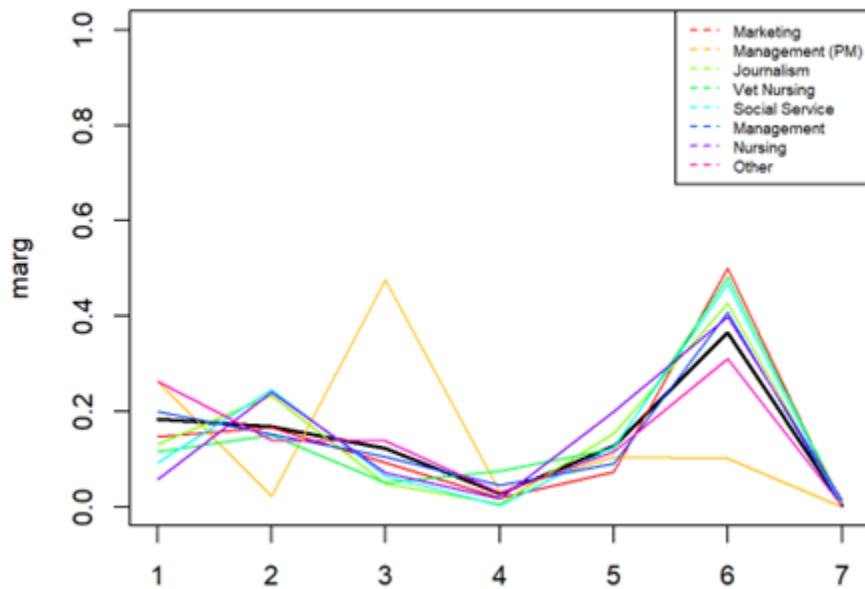
El grup 5 és el que contempla una taxa d'inflació al seu país de procedència més alta. Per altra banda, el 2 és el que menys taxa presenta. La resta de grups són propers a la mitjana global.

Com a conclusió general, sobre els clústers a les variables numèriques, el clúster 1 és el més significatiu, sent el millor a 2 de 6 variables estudiades. A la resta de variables han destacat per igual els grups 2, 3, 5 i 7. Els grups 4 i 6 no han destacat a cap de les variables estudiades.

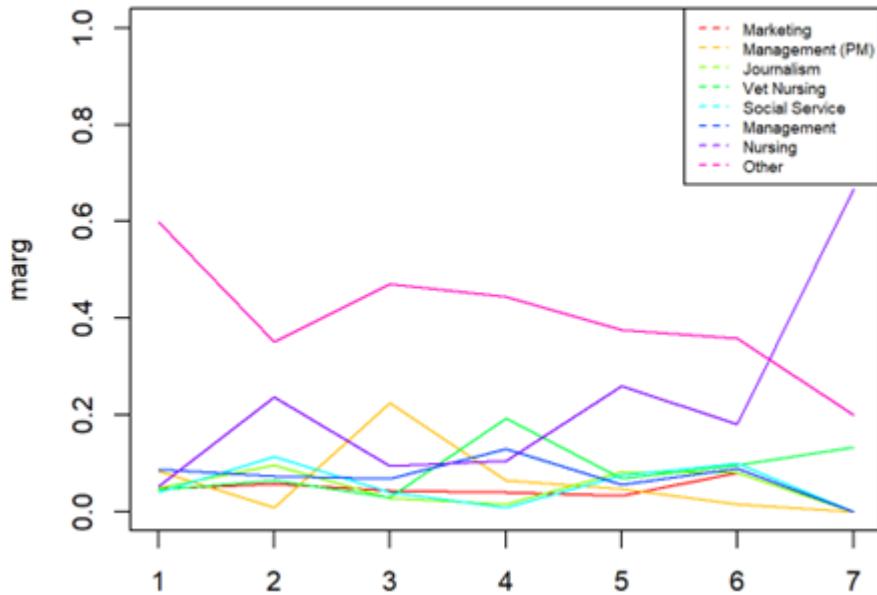
### 7.1.2 Anàlisi gràfica de les variables categòriques

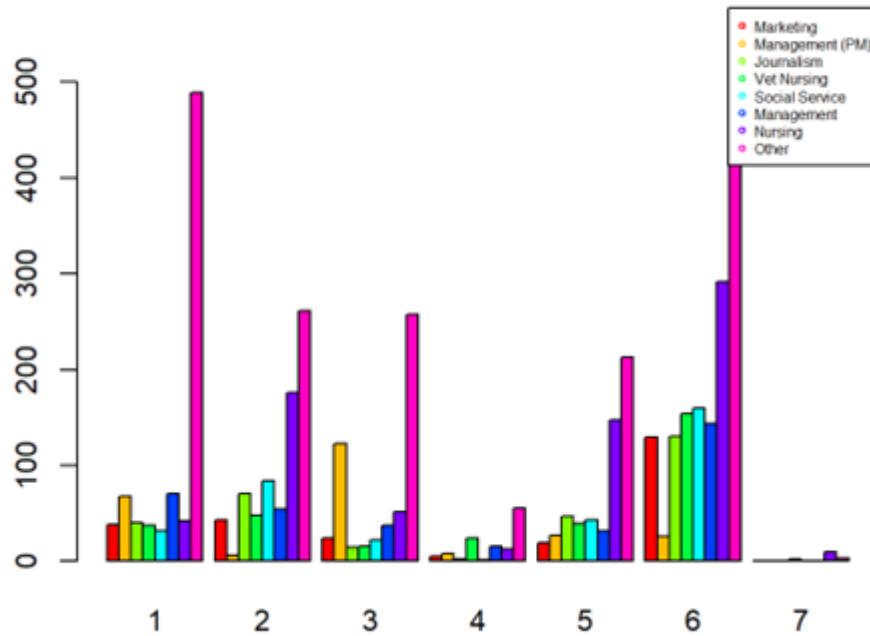
*GRAU DE L'ESTUDIANT*

**Prop. of pos & neg by course**



**Prop. of pos & neg by course**





Als tres gràfics, si un es fixa en les distribucions marginals, en les quantitats a cada clúster i les proporcions per cadascun, es pot observar com:

El clúster 3 és, majoritàriament, dels estudiants de Management (PM).

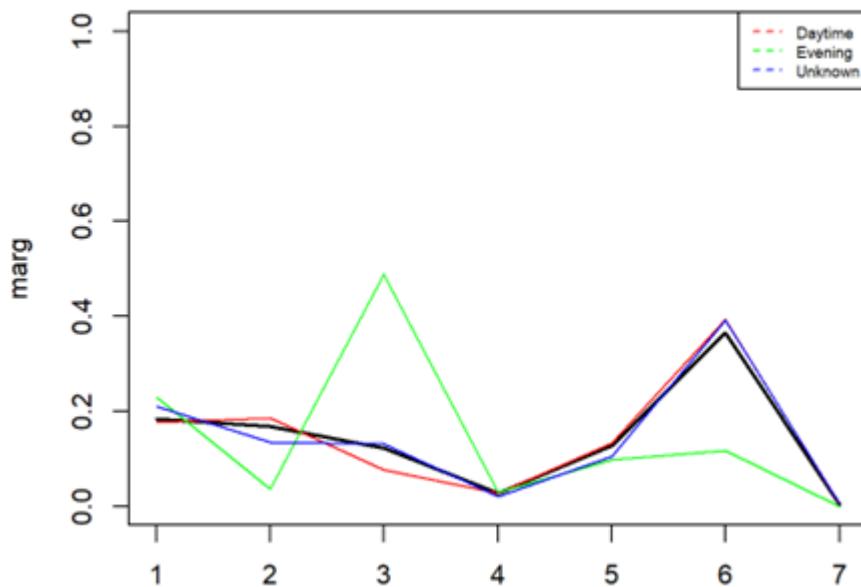
El clúster 7 és dels estudiants d'infermeria.

Al clúster 4, majoritàriament, hi ha els estudiants de veterinària.

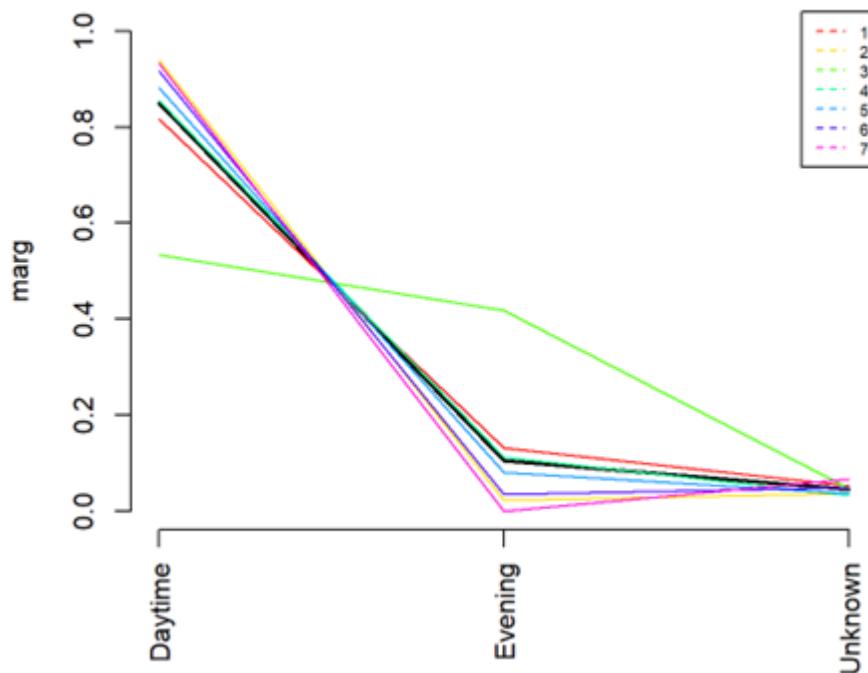
A la resta de clústers no s'observa que cap grau destaquí.

## HORARI DE L'ESTUDIANT

**Prop. of pos & neg by daytime\_attendance**



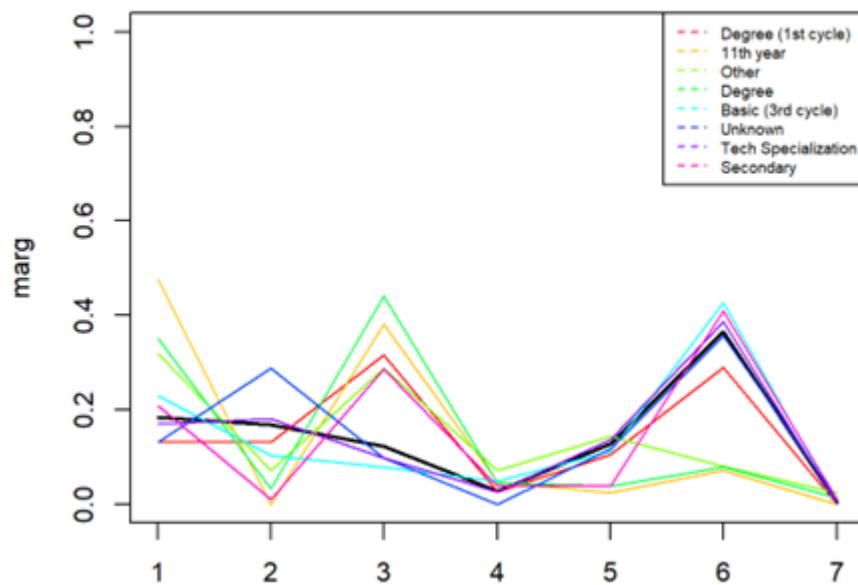
**Prop. of pos & neg by daytime\_attendance**



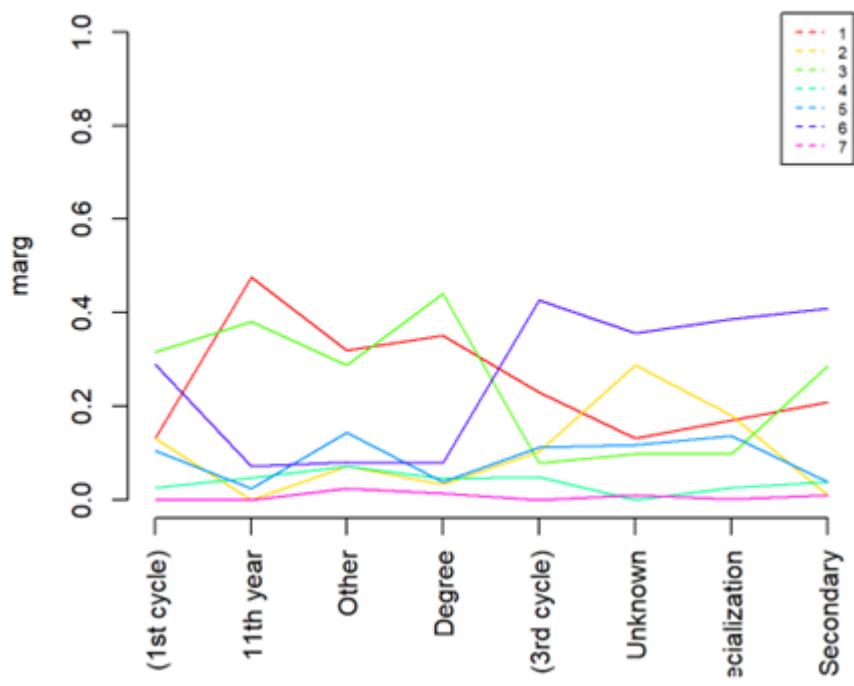
S'observa com quasi tots els clústers segueixen la mateixa distribució. El grup 3 és l'excepció, el qual diu que és significatiu. Es pot dir que és un grup que van tan de matí com de tarda, però, sobretot de tarda. Cal també mencionar que al clúster 6 hi ha menys proporció d'horari de tarda.

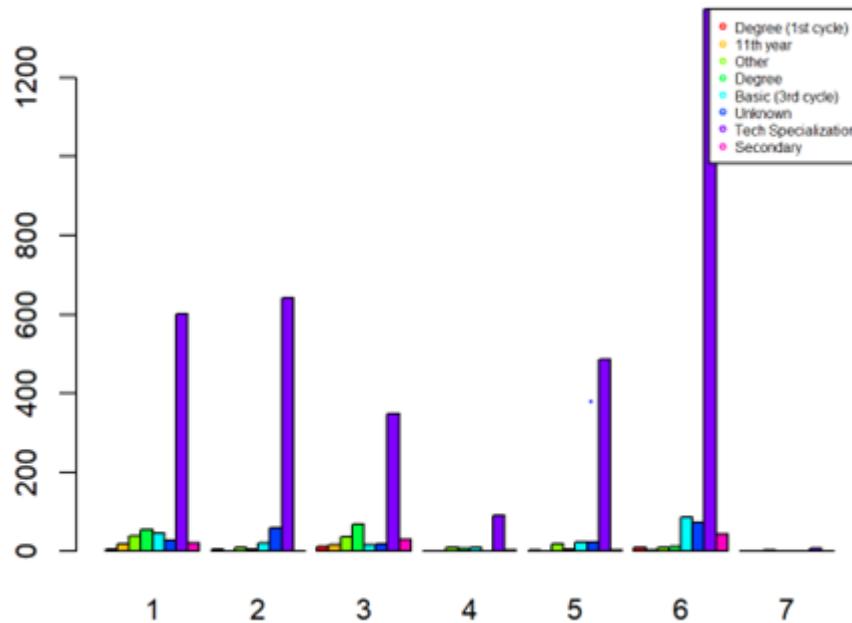
*QUALIFICACIÓ PRÈVIA DE L'ESTUDIANT*

Prop. of pos & neg by previous\_qualification



Prop. of pos & neg by previous\_qualification

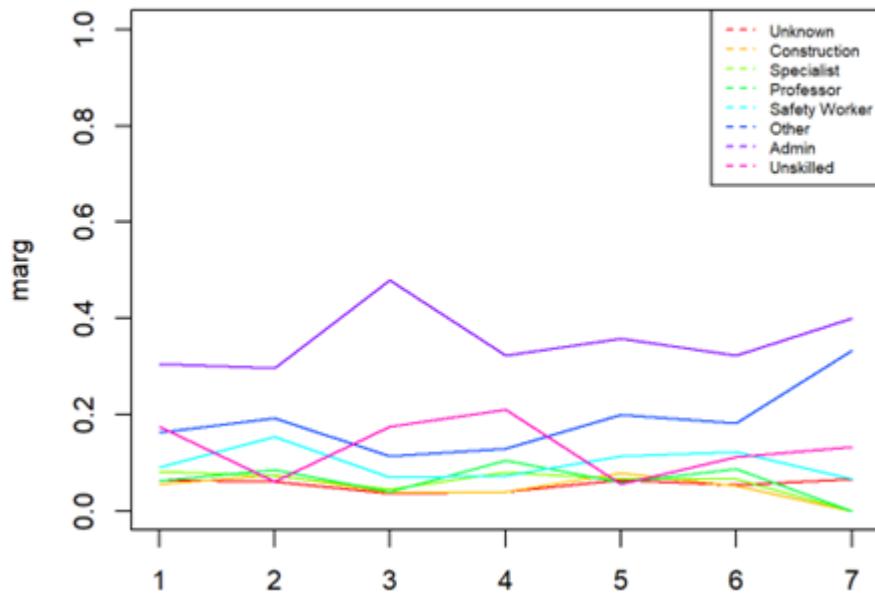




Es veu com la majoria de les titulacions es distribueixen voltant la mitjana global, el que no aporta molta informació, per tal de diferenciar els clústers. Però hi ha algunes diferències, les quals, poden explicar les diferències entre grups. El grup 6 no té quasi estudiants de 11th year, degree o other. El grup 7 és d'estudiants amb titulació prèvia de tech specialization.

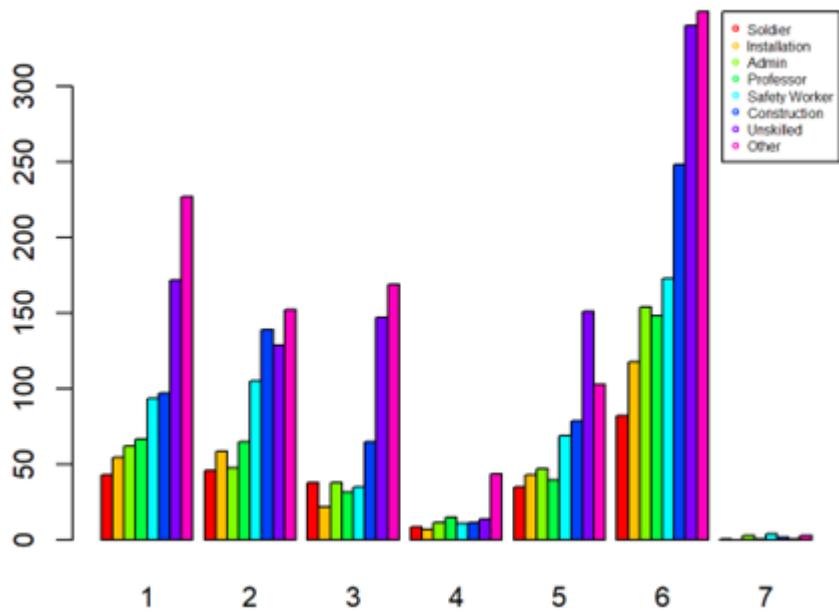
#### OCUPACIÓ DE LA MARE

**Prop. of pos & neg by occupation\_mother**



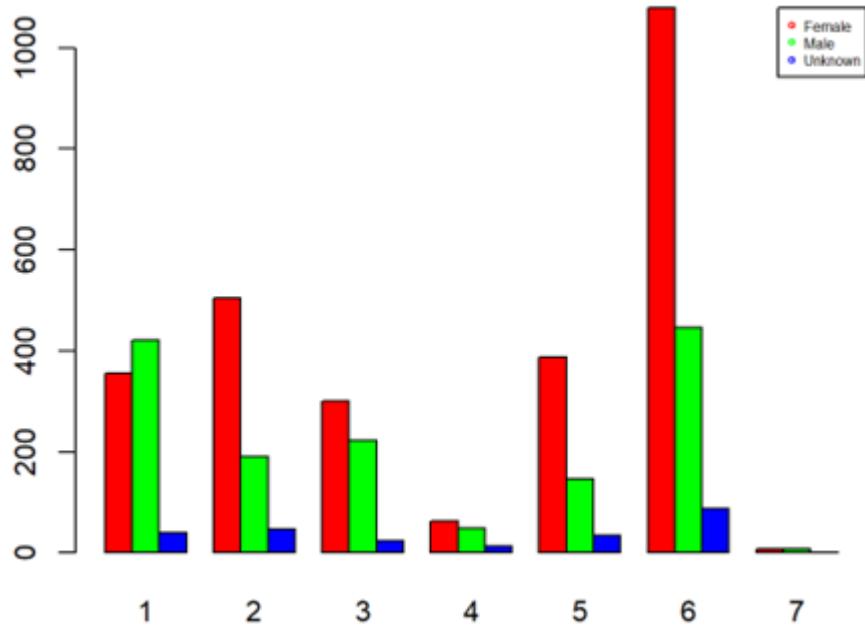
Al gràfic es veu com el clúster 3, majoritàriament, és de les mares Admin. La resta de categories, gairebé, no expliquen els clústers.

## Ocupació del Pare



S'observa com a tots els grups segueix la mateixa distribució. Excepte al clúster 5, majoritàriament, els pares tenen una ocupació unskilled, i al clúster 4 que tenen altre tipus de treball.

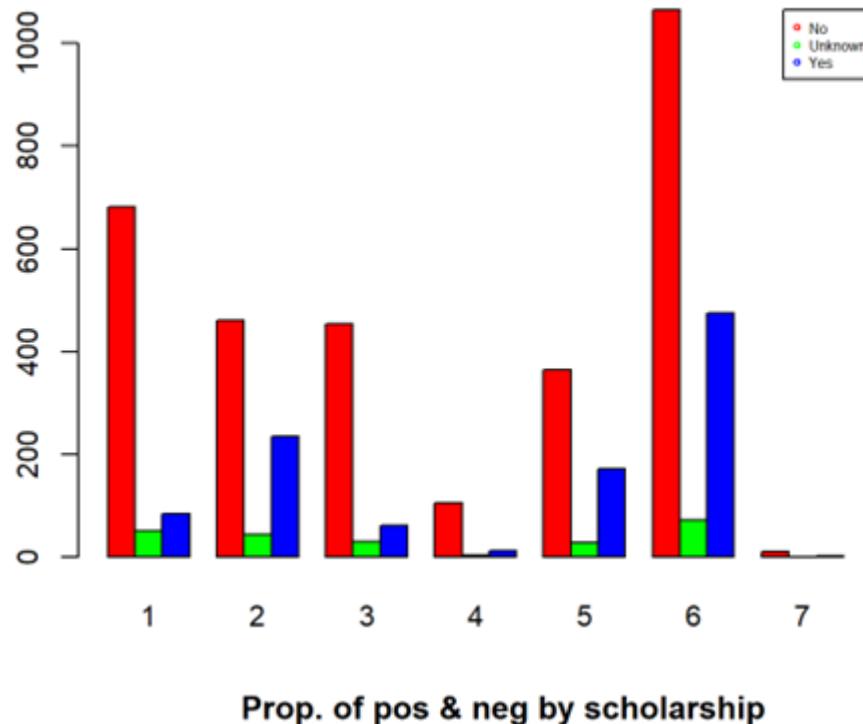
## GÈNERE



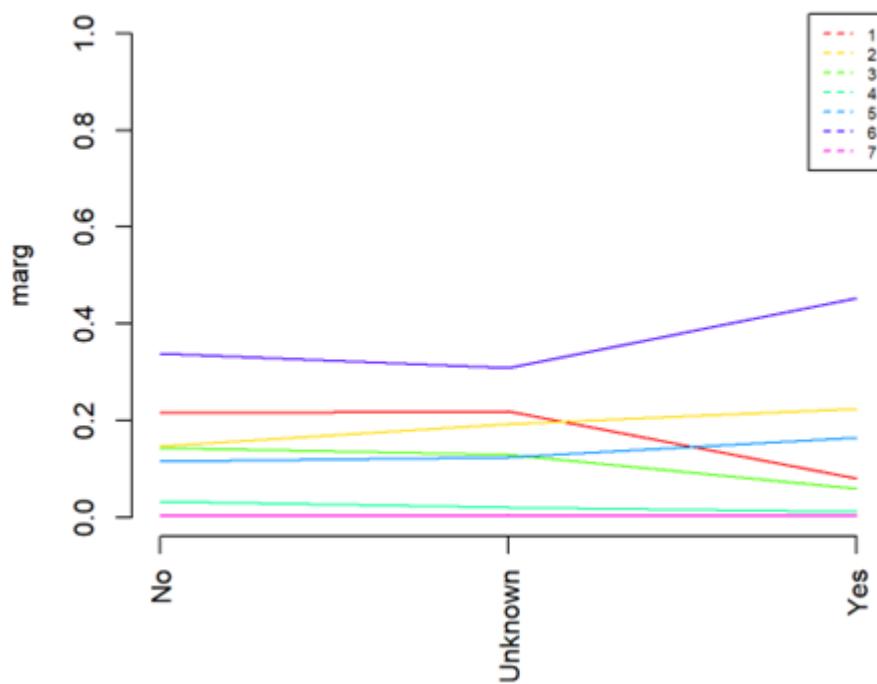
A les distribucions dels clústers, observant el gràfic de barres adossades, es conclou que la majoria dels clústers es distribueixen de la mateixa manera, amb la meitat d'homes, que de dones. El clúster 1 treca aquesta tendència i mostra que el nombre d'homes com de dones és quasi igual, així inclòs, els homes superant a les dones. El grup 4 segueix la mateixa tendència que l'1, i al 6 es veu major quantitat, proporcionalment, de

dones que d'homes, a comparació de com es distribueixen els valors en la resta de clústers.

### BECA



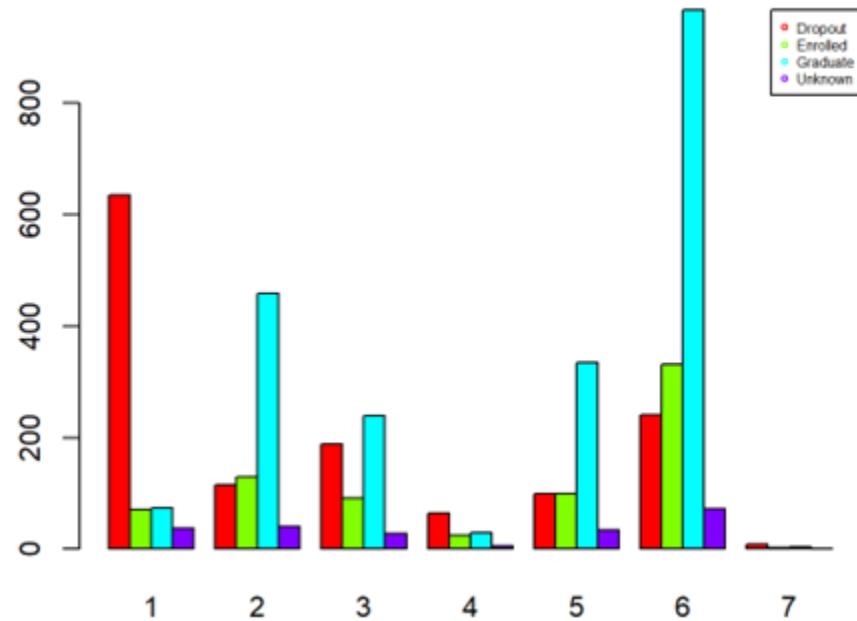
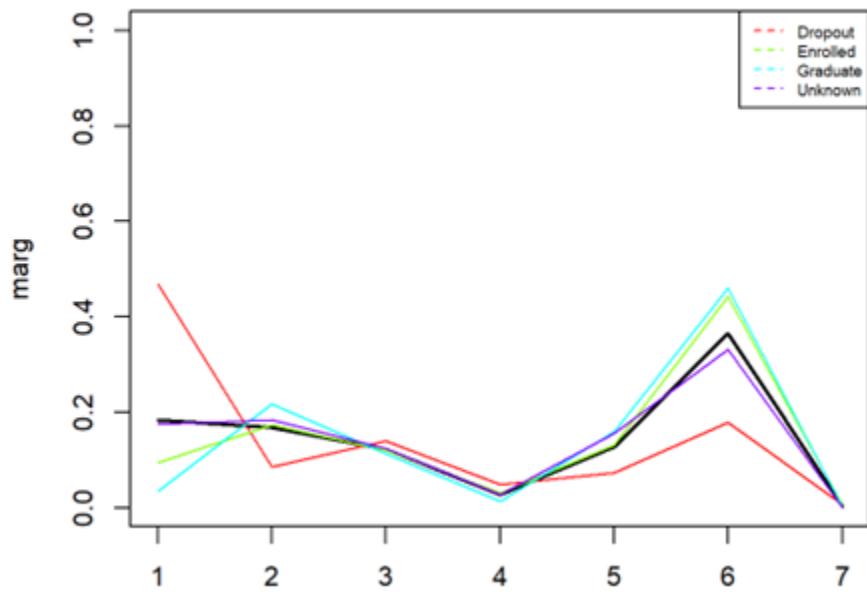
Prop. of pos & neg by scholarship



Més o menys, a tots els grups la proporció de becats i no becats és la mateixa. Excepte al grup 6 on s'observa que hi ha més proporció de becats. El grup 7 no té alumnes becats.

*RESULTAT ACADÈMIC*

**Prop. of pos & neg by target**



Als dos gràfics s'observa com el primer grup està construït pels estudiants que abandonen els estudis, els grups 4 i 7 també mostren un major nombre d'abandonaments, a comparació de la resta de grups. Sobretot al grup 6, i també a la resta, la majoria d'estudiants es graduen.

### 7.1.3 Conclusió sobre els clústers a les variables numèriques i categòriques

**Clúster 1:** Destaca en les notes més baixes als dos semestres, major proporció d'homes. És el grup dels estudiants que deixen d'estudiar i no tenen beca.

**Clúster 2:** Destaca en ser un dels grups més joves, amb la taxa d'inflació al país de procedència més baixa.

**Clúster 3:** Destaca en ser el grup de més edat entre tots, amb mares que treballen d'admin. És el grup dels estudiants de Management, que majoritàriament fan horari de tarda i es graduuen tant com ho deixen.

**Clúster 4:** Destaca en ser el clúster d'estudiants de veterinària, format tant per homes com per dones. És el segon grup que més unitats curriculars suspeses, amb una nota mitjana per sota la mitjana dels grups i majoritàriament deixen d'estudiar.

**Clúster 5:** Destaca en tenir la taxa d'inflació més alta pel país de procedència, amb pares en una ocupació unskilled.

**Clúster 6:** Destaca en ser l'altre grup amb estudiants més joves, majoritàriament dones, amb major proporció d'alumnes becats. És el grup dels estudiants que es graduuen, amb una titulació prèvia bàsica, de secundària o alguna especialització tècnica.

**Clúster 7:** Destaca en ser un grup amb notes molt baixes, suspeses als dos semestres. És el grup dels estudiants d'infermeria, amb major nombre d'unitats curriculars suspeses, que deixen la carrera i amb una titulació prèvia d'especialització tècnica.

### 7.1.4 Comparació dels profilings

S'observa com al primer profiling les variables no prou significatives per a ser estudiades són:

- Nacionalitat de l'estudiant
- Necessitats especials
- Unitats curriculars suspeses al segon semestre
- Taxa de desocupació

Mentre al segon profiling es descarten:

- Nacionalitat de l'estudiant
- Necessitats especials
- Unitats curriculars suspeses al primer semestre
- Taxa de desocupació
- PIB

Al segon profiling s'intercanvien les unitats curriculars suspeses entre semestres i s'afegeix una nova variable a la llista de descarts, el PIB.

Se sap que els dos processos de clustering interpretats presenten algunes diferències notables. En el primer procés, s'han identificat tres grups d'estudiants, mentre que en el segon procés s'han identificat set grups.

En el primer procés, el clúster 1 s'ha identificat com el grup d'estudiants que deixen els seus estudis i estudiants sense beca, i en el segon procés el clúster 1 s'ha identificat gairebé de la mateixa manera. En el segon procés, el perfil d'aquest grup és més detallat. S'observa un altre grup, del segon profiling, que també mostra similituds, aquest és el grup 7, que mostra notes baixes, major nombre d'unitats curriculars

suspeses i a més, són estudiants d'infermeria que deixen la carrera. Un altre grup que es podria apropar a aquest primer grup és el 4, que encara que no té notes tan baixes, és el següent grup amb major nombre d'unitats curriculars suspeses, a més de categoritzar-los com a estudiants de veterinària, i que en parts iguals es graduen i deixen la carrera.

En el primer procés, el clúster 2 s'ha identificat com el grup d'estudiants graduats, amb beca, majoritàriament dones i estudiants de matí. En el segon procés, el clúster 2 s'ha identificat com un dels grups més joves, amb la taxa d'inflació més baixa del país d'origen. Aquestes són dues característiques molt diferents, per la qual cosa sembla que aquests dos clústers són completament diferents. Però, s'observa que el grup 6, creat en el segon procés, mostra bastants similituds, major quantitat de becats, graduats, majoritàriament dones, etc.

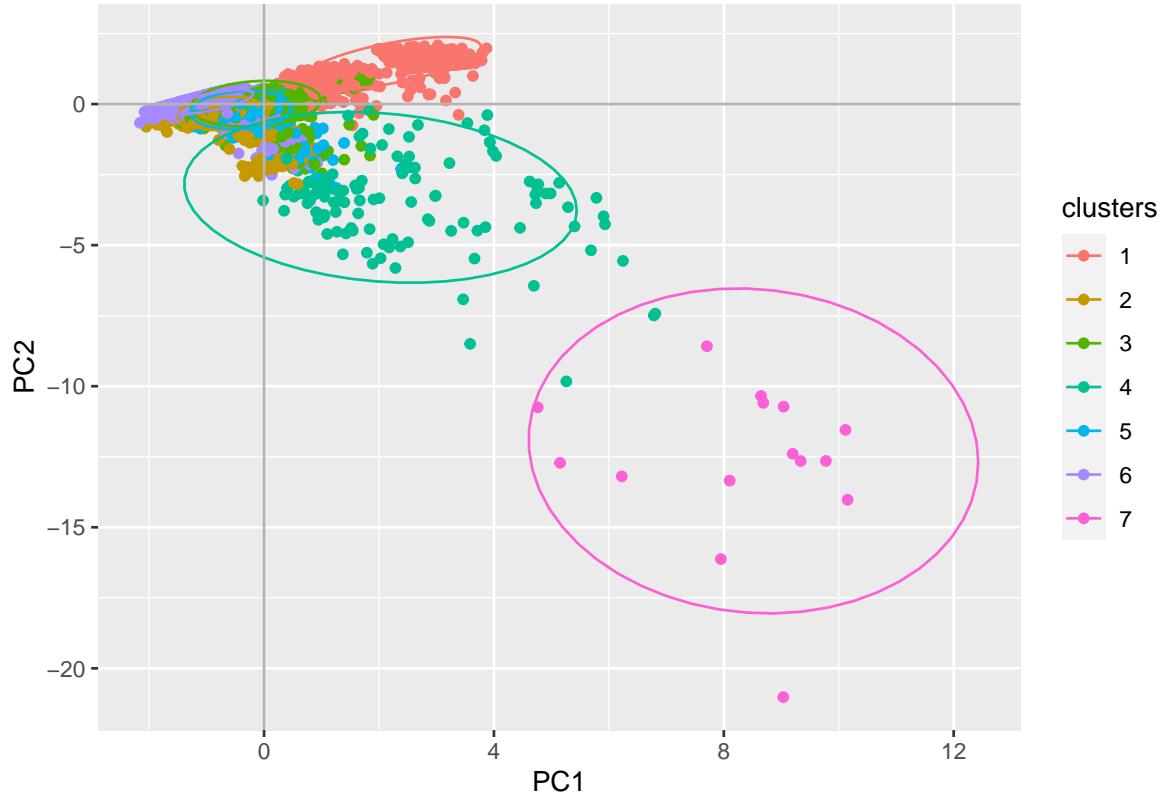
En el primer procés, el clúster 3 s'ha identificat com el grup d'estudiants amb edats més avançades, amb un menor nombre d'unitats curriculars suspeses i un horari de tarda. En el segon procés, el clúster 3 s'ha identificat com el grup d'estudiants més gran, amb mares que treballen com a administratives. A més, són estudiants de Management, que majoritàriament fan horari de tarda i es graduen tant com deixen els seus estudis. Tot i que hi ha algunes similituds, com l'edat avançada i l'horari de tarda, sembla que aquest clúster detalla més característiques del mateix grup.

D'altra banda, tenim el grup 5, del segon procés. Aquest no té prou característiques similars per a observar qualsevol similitud, però per les notes i altres característiques, podria venir del grup 2 del primer profiling.

En general, sembla que el segon procés de clustering és més detallat i precís en la seva descripció dels estudiants i els seus comportaments acadèmics. A més, presenta un major nombre de clústers, cosa que permet una classificació més detallada dels diferents perfils d'estudiants.

## 7.2 Clústers representats sobre els dos primers plans factorials

Es representen les dades dividides per grups en els dos primers plans factorials de l'ACP:



Al gràfic es pot veure que per les dues primeres dimensions, que estan explicades especialment per les notes, els estudiants de veterinària (la segona component només) i els estudiants d'infermeria, els grups que s'allunyen del centre són els 4 i 7.

El grup 7 està conformat per les persones que estudien infermeria, tenen notes baixes, unitats curriculars suspeses i han deixat la carrera. El grup 4 està conformat per estudiants de veterinària que tenen unitats curriculars suspeses. L'únic altre grup que es distingeix una mica és l'1, que són per persones que deixen els estudis. La resta de grups estan barrejats al centre.

## 8 Anàlisi discriminant lineal

L'anàlisi discriminant lineal, o LDA, és un mètode de classificació dins de l'aprenentatge supervisat que permet classificar una variable categòrica entre els seus diferents nivells. Per classificar entre diferents classes s'estimen una o diverses funcions discriminants, que representen hiperplans i separen l'espai d'observacions en diferents regions.

Com que el LDA està basat en mètodes factorials, es prenen només les columnes numèriques com a variables predictives.

La variable que es vol classificar és la variable Target, la qual té 4 modalitats: Dropout, Enrolled, Graduate i Unknown. S'eliminen les observacions de la modalitat Unknown, ja que en realitat són *missing values* i no té sentit que formin una categoria específica en la qual classificar observacions.

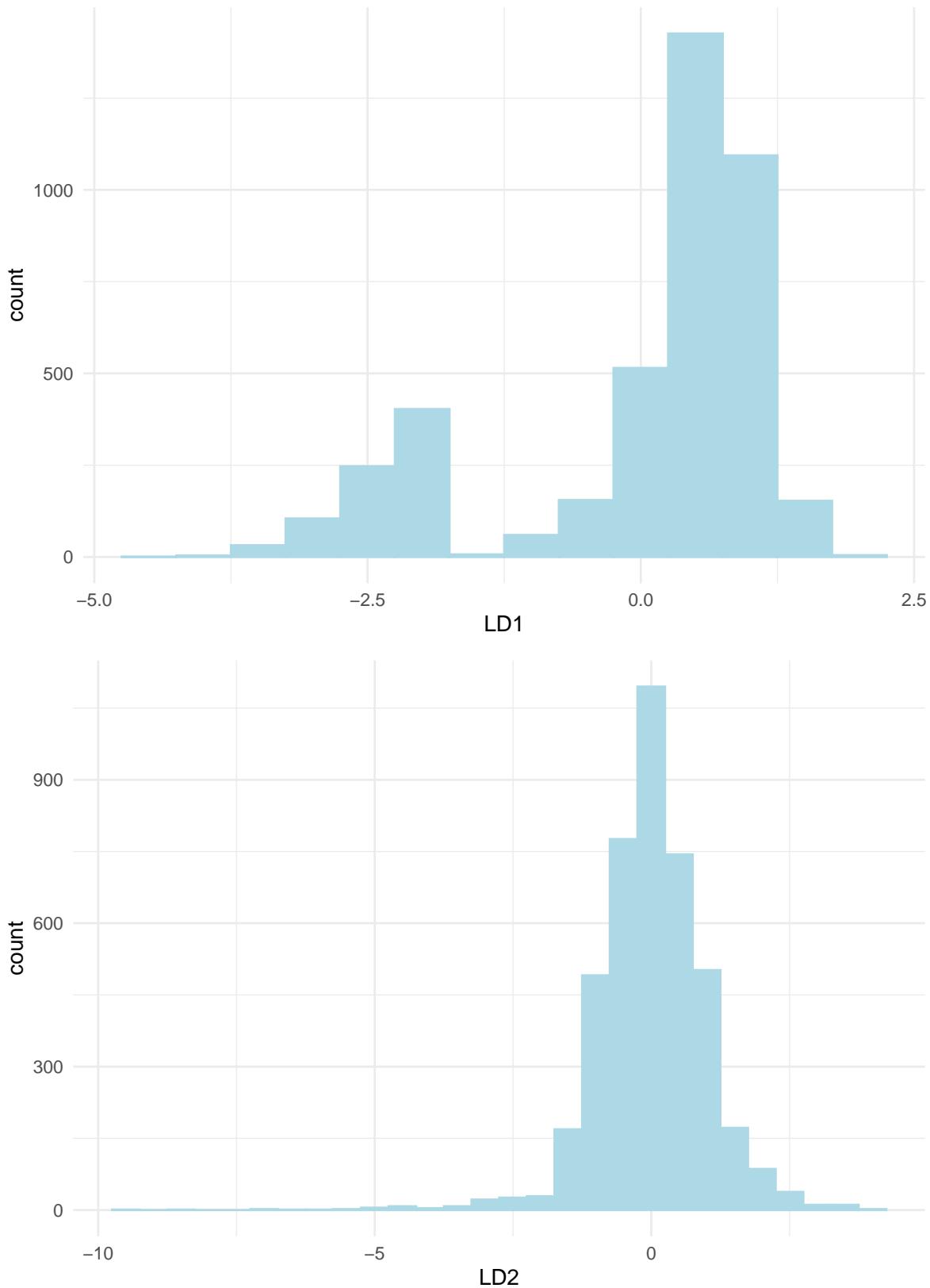
Com que hi ha 3 modalitats per a la variable Target, es busquen dues funcions discriminants per garantir que es pugui arribar a discriminar i separar entre les tres classes. La següent taula mostra els coeficients de les funcions discriminants:

	LD1	LD2
age	-0.0437742	0.0435533
grades_sem_1	0.0242356	0.1867579
no_eval_sem_1	-0.0342150	-0.2431166
grades_sem_2	0.2044008	-0.1323255
no_eval_sem_2	-0.0699950	-0.4638810
unemployment_rate	0.0098169	0.2392478
inflation_rate	-0.0166414	-0.0111436
gdp	0.0054842	0.0252665

Com que s'han calculat dues funcions discriminants, apareixen dues columnes noves que s'afegeixen a la base de dades que conté les columnes numèriques de la base original.

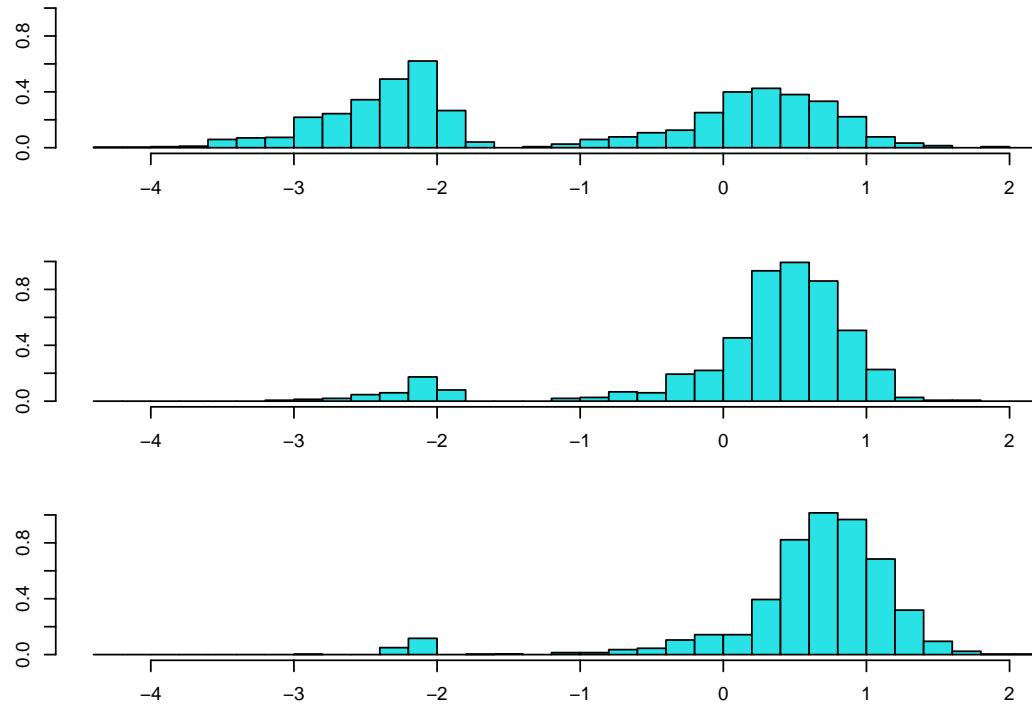
```
## [1] "variable LD1 Vw = 0.999999999999999 Vb = 1139.70636146981 separation = 1139.70636146981"
## [1] "variable LD2 Vw = 0.999999999999995 Vb = 10.6790279023928 separation = 10.6790279023929"
```

Es dibuixen en un histograma els coeficients de cada funció discriminant.

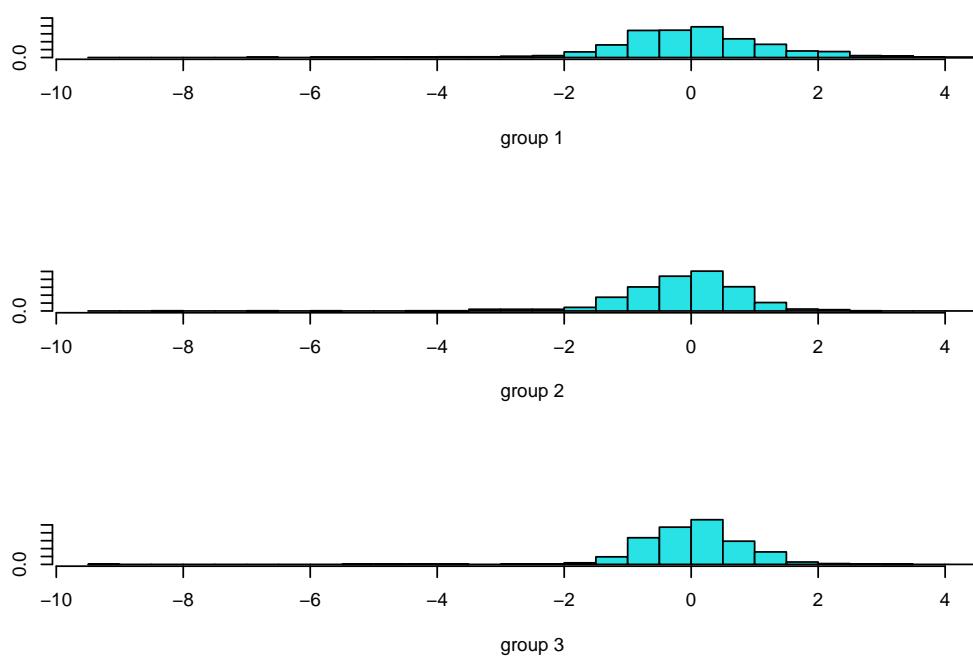


Es poden també representar en histogrames els coeficients de la funció discriminant segons el nivell de la variable Target.

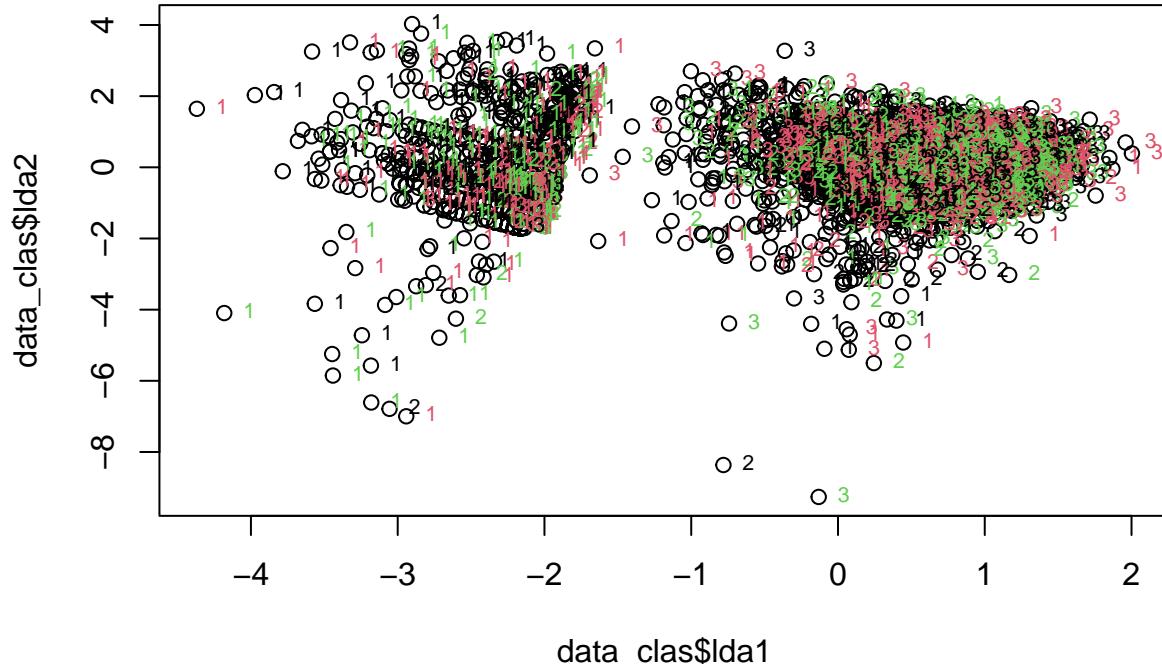
LD1



LD2



Es poden també representar les diferents classes de la següent manera:



```
##          age      grades_sem_1      no_eval_sem_1      grades_sem_2
## 0.04355331     0.18675786    -0.24311662    -0.13232545
## no_eval_sem_2 unemployment_rate      inflation_rate           gdp
## -0.46388104     0.23924779    -0.01114359     0.02526652
##          age      grades_sem_1      no_eval_sem_1      grades_sem_2
## -0.043774186    0.024235650   -0.034215037   0.204400831
## no_eval_sem_2 unemployment_rate      inflation_rate           gdp
## -0.069995008    0.009816912   -0.016641377   0.005484226
```

En tots els gràfics s'aprecia el mateix, no es discriminen bé les modalitats.

Un cop s'han predit les classes de la variable Target, es pot calcular la matriu de confusió per avaluar la precisió de les funcions discriminants que s'han calculat a l'hora de separar les classes.

```
##
##    1     2     3
## 1353  750 2104
```

S'observa, però, que la precisió del model és força baixa, al voltant del 15%.

```
## [1] 0.1568814
```

Amb la precisió calculada, es pot obtenir l'error de classificació:

```
## [1] 0.8431186
```

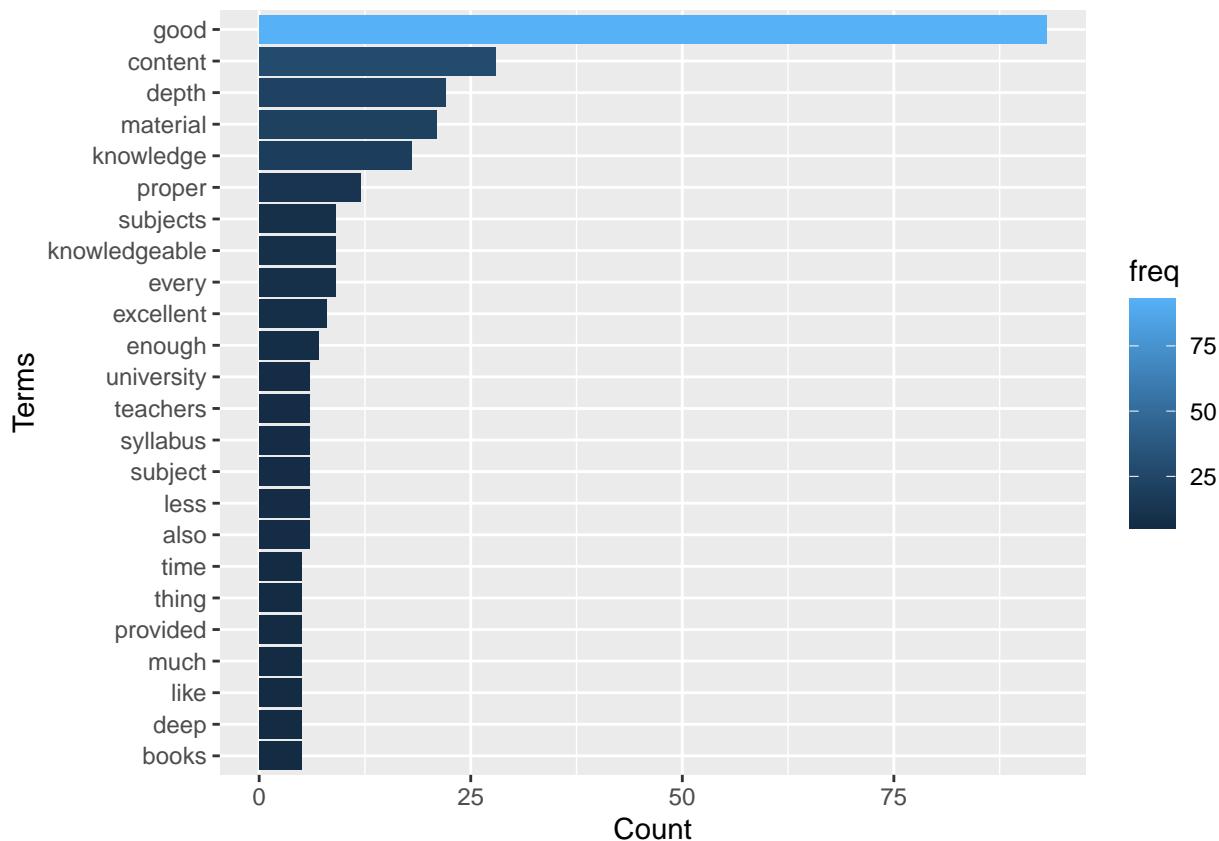
## 9 Anàlisi textual

### 9.1 Preprocessament

Per poder desenvolupar l'anàlisi textual primer cal preprocessar la base de dades amb l'objectiu de mantenir només aquella informació rellevant que permeti una anàlisi més precisa.

Aquest preprocessament inclou eliminar els espais en blanc, les majúscules, els signes de puntuació, els números, altres caràcters especials, les stop words (to, the, a...), etc. A més, s'aplica el procés de lematització de les paraules (transformar les paraules a la seva arrel per poder trobar més relacions entre textos diferents).

S'agafen les paraules que surten 5 vegades o més per tenir aquella informació rellevant i les representem gràficament:

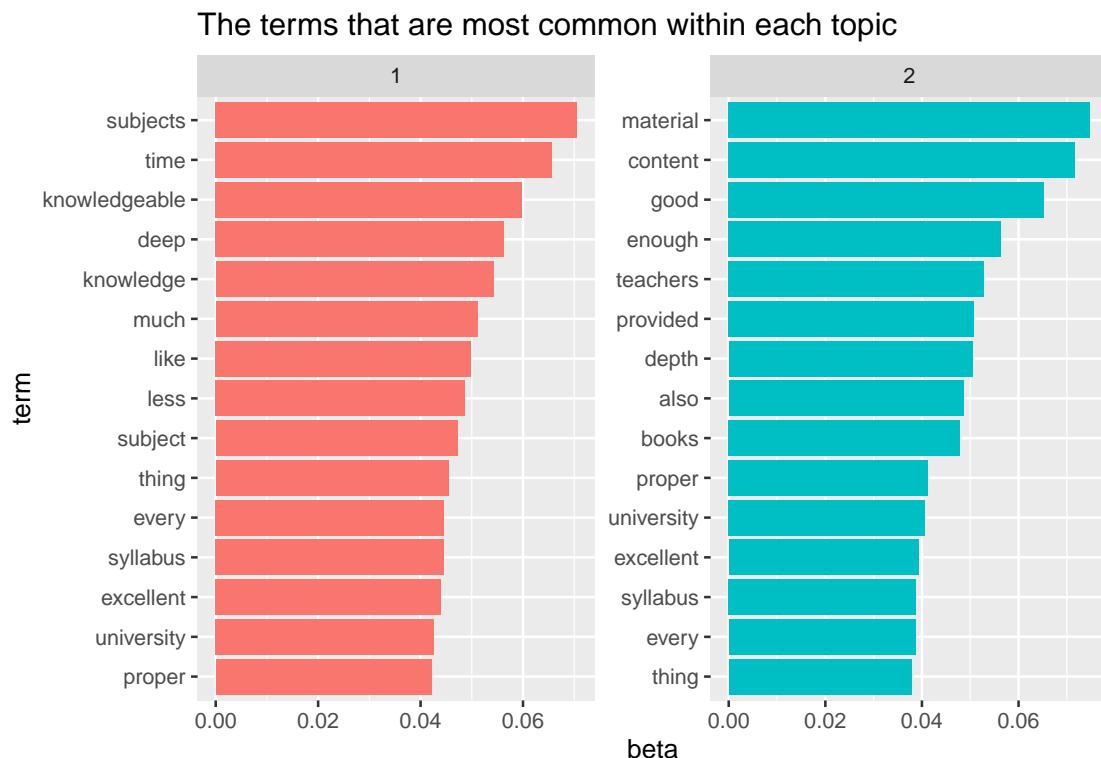




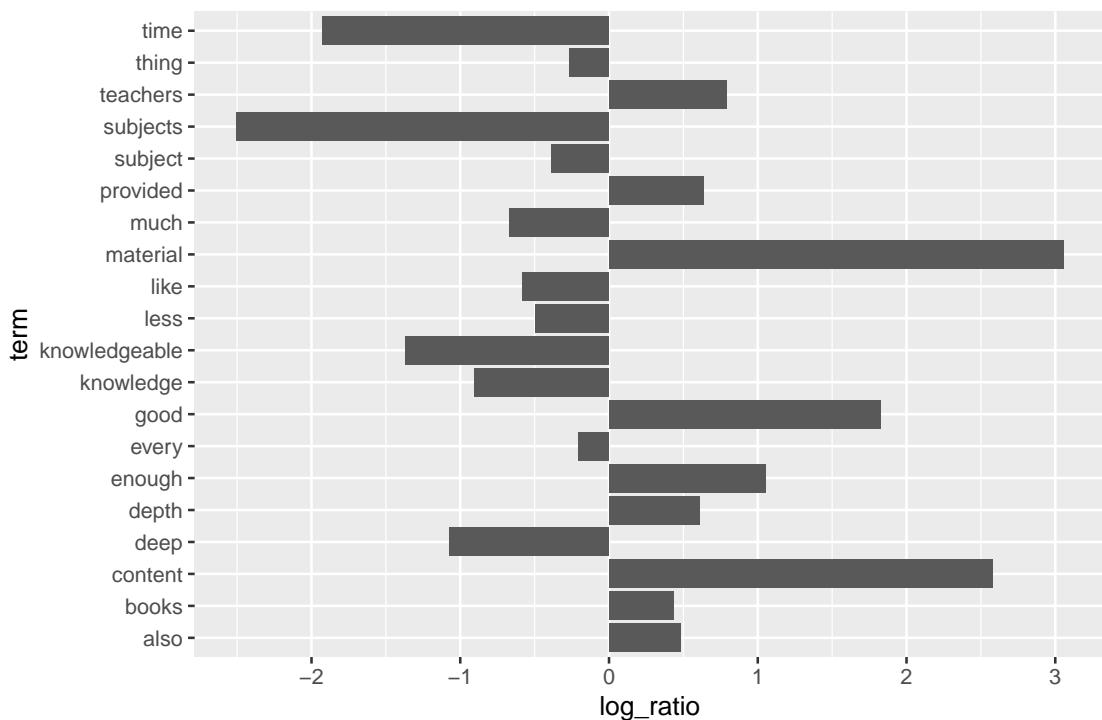
## 9.2 Topic modeling

### 9.2.1 Per dos tòpics

Dividint les respostes dels alumnes en dos temes:



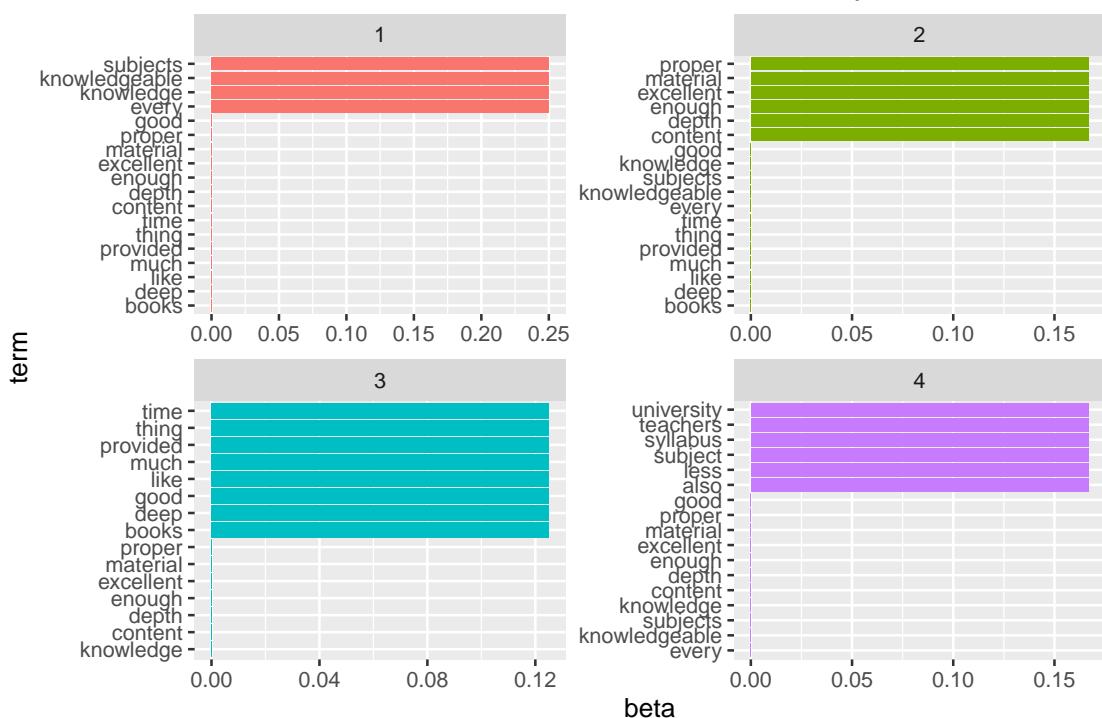
### Terms with the great difference in beta between two topics



#### 9.2.2 Per quatre tòpics

Dividint les respostes dels alumnes en quatre grups:

### The terms that are most common within each topic



En el gràfic anterior es pot observar que el primer tema tracta els coneixements obtinguts a l'assignatura. El segon engloba els continguts i els materials del curs. El tercer està format sobretot per sentiments i opinions dels alumnes, mentre que el quart tema són la universitat i els professors.

Cal tenir en compte que hi ha un nombre limitat de respostes i aquestes són poc diverses, el que implica que l'anàlisi de temes no és del tot significatiu.

## 10 Anàlisi comparativa i conclusió

Es fan servir mètodes per crear subpoblacions i poder diferenciar perfils dins les dades. La intenció és reduir la dimensionalitat de la matriu original mantenint la màxima informació possible. Es busca tenir un 80% d'inèrcia acumulada, per obtenir-la calen: 5 components principals a l'ACP, i 15 dimensions a l'ACM.

Al clustering jeràrquic, en primer lloc, s'arriba a la conclusió que el nombre òptim de clústers és 3, a partir de diversos mètodes: el de Ward amb la dissimilitud de Gower al quadrat, ajudant-se d'un dendrograma, el coeficient de Silhouette i la funció de R nbclust(). Fent-ne el profiling, es caracteritza el clúster 1 com al de les notes més baixes i més suspensos, que són els que deixen els estudis o no tenen beca. El segon clúster és el de les notes més altes i persones de països amb PIB més alt, són estudiants que s'han graduat, amb beca, majoritàriament dones i que tenen a classe al matí. El tercer clúster és el de les edats més altes i menys unitats suspeses, són persones que atenen de tarda i que tenen probabilitat igual de graduar-se com d'abandonar.

A l'anàlisi per components principals s'ha extret que, per arribar al 80% d'inèrcia, el nombre de components principals necessaris són cinc. Expliquen, en concret, el 84.32% de la inèrcia. Per tant, permeten explicar la major part de la variabilitat de les dades, el que significa que n'estreuen quasi tota la informació. En concret, s'ha trobat que per a la primera component les variables que més contribueixen són l'edat, de forma directa, i les notes dels dos semestres, de forma inversa. Veient les categòriques, es pot observar que aquest primer component es relaciona amb graduar-se o abandonar, els valors negatius correlacionats amb graduar-se i els positius amb abandonar. Les altres components no donen informació d'una forma tan directa.

Quan es fa el clustering a partir d'aquesta anàlisi, l'ACP, el nombre de clústers que en resulten són 7, que estan relacionats amb els originals, tot i que els especifiquen i diversifiquen més.

A l'anàlisi de correspondència múltiple, la conclusió és que, en funció de les zones dels gràfics, es poden distingir certes modalitats de diverses variables, el que implicaria associació. No es pot, però, arribar a trobar la relació entre les variables, i el percentatge de variància explicat per aquesta anàlisi és baix, el 24.25%. Les associacions principals trobades als diversos quadrants defineixen tres perfils, que resulten bastant específics. Es deuen repetir diverses vegades, però no semblen ser representatius d'una part significativa de l'alumnat.

A partir d'aquests mètodes s'han trobat resultats que, malgrat no coincidir de forma exacta, tenen sentit els uns amb els altres. Es defineixen diversos perfils amb diferents probabilitats de graduar-se i diferents notes, que són coherents entre els diversos mètodes, tot i que cadascun els presenta d'una forma diferent i amb un grau més alt o més baix d'especificitat.

A l'anàlisi discriminant lineal s'han buscat dues funcions discriminants per separar les tres modalitats de la variable target: enrolled, drop-out i graduate. S'ha observat que no és fàcil discriminar les tres modalitats a partir de les variables utilitzades. La precisió del model és baixa, del 15.7%, és a dir, que l'error de classificació és del 84.3%.

Es conclou, per tant, que la probabilitat de graduar-se no és independent de diversos factors com ara l'edat, el gènere, tenir beca o no, el PIB del país de procedència i l'horari en què es va a classe, ja que si es va de tardes de forma voluntària, és més probable que sigui perquè la persona treballa al matí, el que vol dir que té menys temps a dedicar als estudis, entre altres. També s'observa que hi ha diferents perfils d'estudiant, que tenen més o menys probabilitat de graduar-se en funció dels factors esmentats. Tot i això, cal tenir en compte que l'anàlisi discriminant mostra que no és possible predir de forma precisa la probabilitat de graduar-se d'una persona a partir d'aquests factors, així que, malgrat que es pot establir una correlació, aquesta no és prou forta per a fer prediccions.

L'anàlisi textual s'ha realitzat amb dades alienes a les originals, en vista que aquestes no tenien cap variable que apliqués. També tracten l'àmbit acadèmic. Després del preprocessament, s'ha observat que la paraula més comuna és good, i que hi ha quatre tòpics principals que són: coneixements obtinguts a les assignatures, continguts i materials del curs, sentiments i opinions de l'alumnat, i la universitat i el professorat. La base de dades no té una mida gaire gran, així que els resultats no són especialment significatius.

## 11 Pla de treball real

En finalitzar el treball es compara el pla de treball inicial amb el ritme de treball real. Per això, en aquest apartat es revisa punt a punt el pla de treball proposat al principi, per detectar possibles canvis i la incidència de riscos.

### 11.1 Diagrama de Gantt

Per tal de complir els objectius del treball i seguir tots els punts que consten en el guió, es va dividir la feina en les següents tasques a desenvolupar.

1. Motivació del treball, descripció formal de les dades i profiling de les dades crues
2. Preprocessament
3. Profiling dades preprocessades i clustering jeràrquic
4. Preparar informe D3 i presentació 1
5. ACP i ACM
6. Clustering jeràrquic de l'ACP i l'ACM
7. Anàlisi comparativa
8. Conclusions
9. Pla de treball real
10. Preparar informe D4 i presentació 2

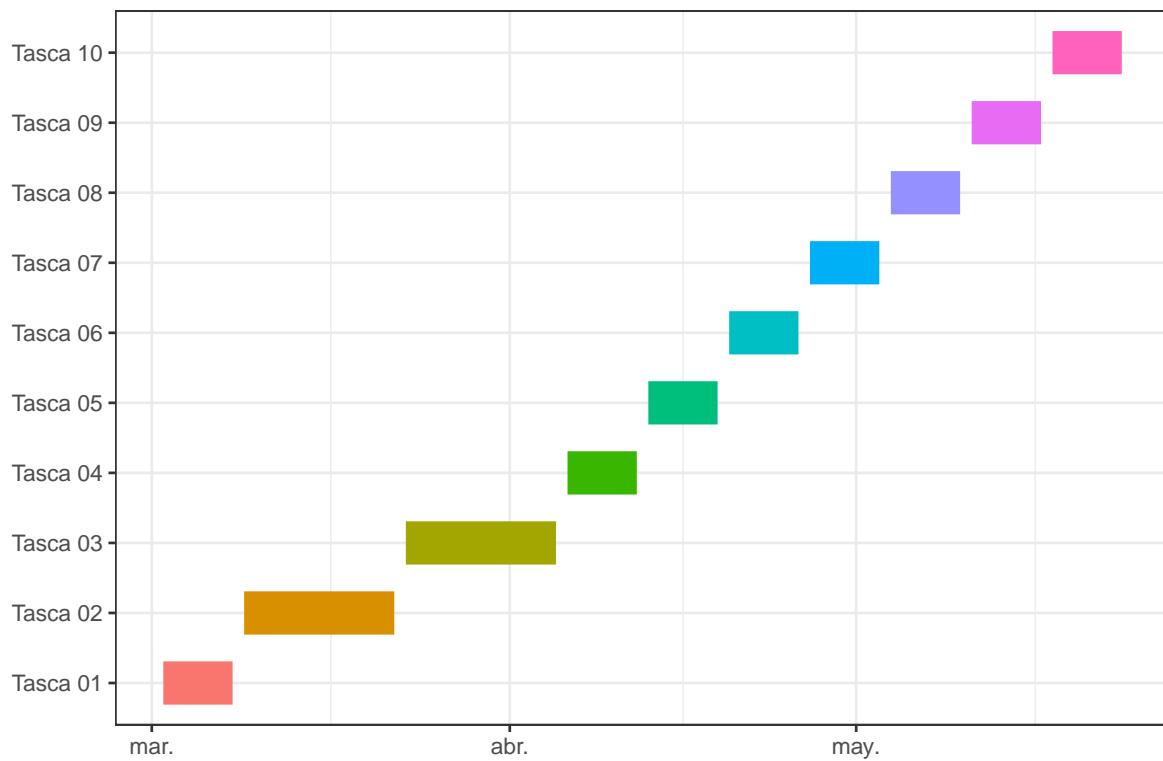
Durant les últimes setmanes es van afegir dos apartats més per consideració dels professors: anàlisi discriminant i anàlisi textual. Llavors, les tasques han quedat així:

...

7. Anàlisi discriminant
8. Anàlisi textual
9. Anàlisi comparativa
10. Conclusions
11. Pla de treball real
12. Preparar informe D4 i presentació 2

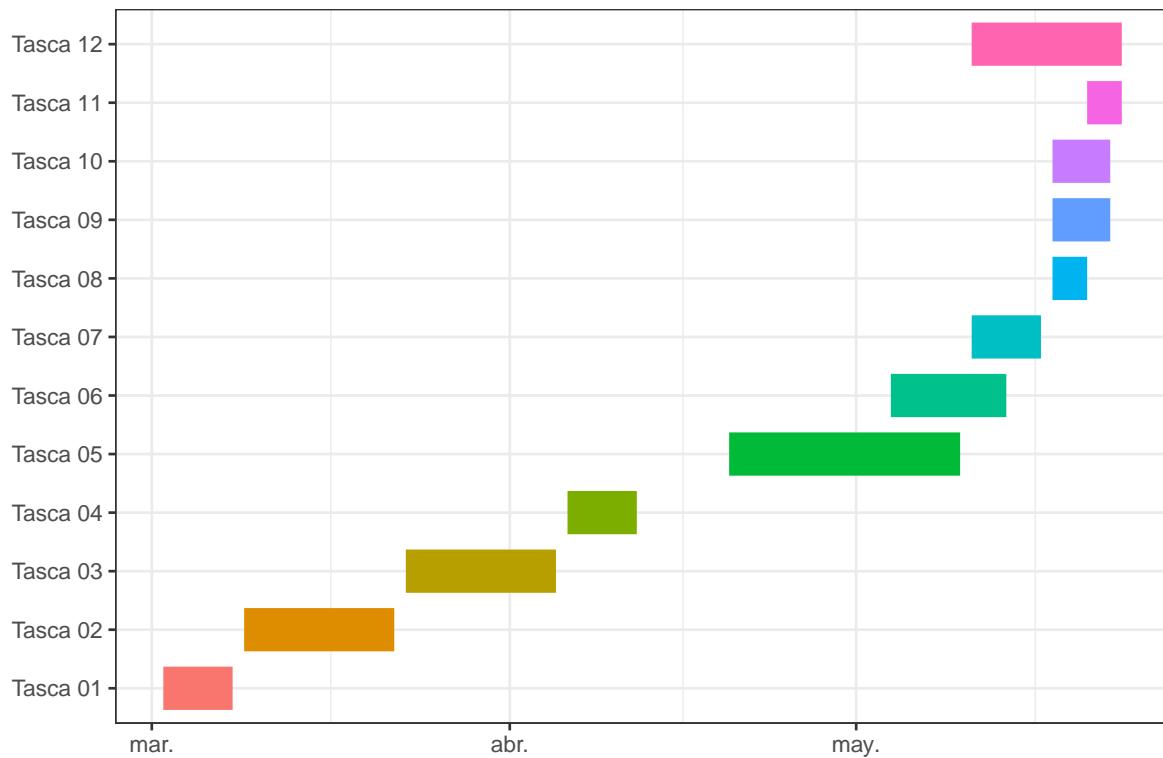
Aquesta és la planificació de tasques proposada a l'inici del treball.

### Diagrama de Gantt



I aquesta és la temporalització i el ritme real de les tasques.

### Diagrama de Gantt real



El pla de treball real s'ha vist afectat, especialment en les etapes finals, pel ritme de les classes de teoria, ja que es van plantejar tasques per abans que s'expliqués el tema corresponent a classe. La tasca de l'ACP i l'ACM n'és un exemple. A més, es va sobreestimar el temps que requeririen algunes tasques, i se'n van subestimar d'altres.

Al maig, s'han encavalcat moltes tasques a causa de dos factors: les dues tasques extres no previstes i la subestimació del temps necessari per a l'ACP i l'ACM, així com al clustering jeràrquic d'aquests.

## 11.2 Assignació de tasques

Les tasques descrites en l'anterior secció es van repartir entre els integrants del grup de la següent manera:

Tasca	Aina i Oscar	Lucía i Laura T.	Mireia i Laura V.	Berta i Albert
Motivació treball	X			
Descripció formal dades		X		
Profiling dades crues			X	
Preprocessament				X
Profiling dades preprocessades		X		
Clustering jeràrquic	X			
Informe D3 i presentació 1	X	X	X	X
ACP			X	
ACM				X
Clustering jeràrquic ACP i ACM	X			
Anàlisi comparativa		X		
Conclusions	X	X	X	X
Pla treball real				X
Informe D4 i presentació 2	X	X	X	X

Tots els grups han dut a terme les tasques assignades. Les dues tasques extres es van assignar als grups amb més disponibilitat de temps en el moment en què es van proposar els nous apartats. L'assignació de tasques final ha estat:

Tasca	Aina i Oscar	Lucía i Laura T.	Mireia i Laura V.	Berta i Albert
Motivació treball	X			
Descripció formal dades		X		
Profiling dades crues			X	
Preprocessament				X
Profiling dades preprocessades		X		
Clustering jeràrquic	X			
Informe D3 i presentació 1	X	X	X	X
ACP			X	
ACM				X
Clustering jeràrquic ACP i ACM	X			
<b>Anàlisi discriminant</b>			X	
<b>Anàlisi textual</b>				X
Anàlisi comparativa		X		
<b>Conclusions</b>		X		
Pla treball real				X
Informe D4 i presentació 2	X	X	X	X

### 11.3 Pla de riscos

Amb la premissa de prevenir que qualsevol imprevist s'agreugi i poder continuar la planificació, es va dur a terme el següent pla de riscos:

Risc	Impacte	Solució
Falta de temps acadèmic, retard en les tasques	Alt	Planificació amb suficient antelació. Si cal, reajustar el calendari, allargar i escurçar terminis.
Pèrdua de documents	Alt	Possible recuperació de versions anteriors. Tenir còpies de totes les actualitzacions (Drive i Github).
Renúncia o pèrdua de l'avaluació contínua	Moderat	Es reorganitzen els grups.
Incapacitat per treballar (indisposició, malalties, problemes personals, problemes informàtics, etc.)	Moderat	Si la tasca té poc pes, la fa la parella. Si és molta feina, s'encarrega un altre grup que no tingui cap tasca en el moment.
Conflictos entre parelles	Baix	Intentar solucionar els problemes amb ajuda d'algun membre com a mediador. Si no s'aconsegueix, reestructuració dels grups.

Al final del projecte, s'han revisat la incidència dels riscos previstos i de no previstos. Queden resumits en la taula següent:

Risc	Incidència	Accions
Falta de temps acadèmic, retard en les tasques	Alt	S'ha reajustat el calendari, allargant i escurçant terminis.
Pèrdua de documents	No succeït	
Renúncia o pèrdua de l'avaluació contínua	No succeït	
Incapacitat per treballar (indisposició, malalties, problemes personals, problemes informàtics, etc.)	No succeït	
Conflictes entre parelles	No succeït	
<b>Calendari no compassat amb les classes teòriques</b>	<b>Moderat</b>	<b>S'ha reorganitzat el calendari, començant les tasques tan aviat com s'ha pogut</b>