Tipologia-PRAC02

GALVEZ BAIXENCH, CONCEPCIÓN i VERA NIETO, JUAN ANTONIO

2025-05-27

COMENTARIS PREVIS A LA RESOLUCIÓ DE LA PRÀCTICA

El dataset escollit per a realitzar aquesta segona pràctica no és el resultat de la primera pràctica, com es demanava en aquest enunciat. Això és degut a que quan es va realitzar la primera pràctica no es va tenir en compte que el dataset resultant hauria de tenir certes característiques que el permetéssin fer-ho servir en aquesta pràctica com la neteja de les dades, la creació de diferents models d'aprenentatge (tant supervisats com no supervisats), etc. Per això, i un cop concertat a través del fòrum de l'assignatura amb la professora la possibilitat de fer-ne servir d'altres datasets diferents, hem fet servir el següent conjunt de dades per poder realitzar completament tots els exercicis d'aquesta pràctica:

https://www.askamanager.org/2021/04/how-much-money-do-you-make-4.html

https://docs.google.com/spreadsheets/d/1IPS5dBSGtwYVbjsfbaMCYIWnOuRmJcbequohNxCyGVw/edit?resourcekey=&pli=1&gid=1625408792#gid=1625408792

Pegunta 1.Descripció del dataset.

1. Perquè és important i quina pregunta/problema pretén respondre?
2. Resumeix breument les variables que el formen i el seu tamany.

RESPOSTA:

Descripció del dataset

Aquest dataset conté respostes d'una enquesta salarial amb 28.141 registres i 18 variables. Les dades han estat recollides d'un formulari en línia disponible a *https://www.askamanager.org/2023/04/how-much-money-do-you-make-6.html* i representen informació relacionada amb la situació laboral, experiència, educació i compensació econòmica de professionals arreu del món.

L'enquesta recull informació a través de tres tipus de preguntes:

1. Preguntes de selecció única.
2. Preguntes de selecció múltiple
3. Preguntes de text lliure

No totes les preguntes son obligatòries de contestar.

Importància del dataset Aquest conjunt de dades és important perquè permet:

1. Analitzar diferències salarials segons sector, gènere, ubicació geogràfica, experiència i educació.
2. Identificar desigualtats o tendències en el mercat laboral.

Pregunta central que pretén respondre:

Quins factors influeixen en la variació dels salaris entre professionals de diferents perfils, gènere, raça i educació?

Resum de les variables

Variable original	Descripció
Timestamp	Data i hora de la resposta
How old are you?	Franja d'edat
What industry do you work in?	Sector laboral
Job title	Títol del lloc de treball
If your job title needs additional context	Informació opcional sobre el lloc que ocupa
What is your annual salary?	Salari anual
How much additional monetary compensation	Compensació addicional (bonus, hores extra, etc.)
Please indicate the currency	Moneda
Other currency	Especificació si es marca "Other" a l'anterior
Income context	Context extra sobre el salari (opcional)
What country do you work in?	País de treball
US State	Estat nord-americà (si aplica)
City	Ciutat de treball
Years experience in field	Anys d'experiència en el camp específic
Years experience overall	Anys d'experiència total
Education level	Nivell d'educació
Gender	Gènere
Race	Raça/etnicitat

OBS: En aquest informe només es mostrar el codi i el resultat necessari de les visualitzacions més importants, deixant la resta de comentaris de com s'ha fet dins del Rmarkdown disponible al GitHub del projecte.

Un cop llegit l'arxiu. mirarem el nom de les columnes i dimensions del dataset carregat.

```
## [1] 28141    18
```

```
##  [1] "Timestamp"
##  [2] "How old are you?"
##  [3] "What industry do you work in?"
##  [4] "Job title"
##  [5] "If your job title needs additional context, please clarify here:"
##  [6] "What is your annual salary? (You'll indicate the currency in a later question. If you are part
##  [7] "How much additional monetary compensation do you get, if any (for example, bonuses or overtime
##  [8] "Please indicate the currency"
##  [9] "If \"Other,\" please indicate the currency here:"
## [10] "If your income needs additional context, please provide it here:"
## [11] "What country do you work in?"
## [12] "If you're in the U.S., what state do you work in?"
## [13] "What city do you work in?"
## [14] "How many years of professional work experience do you have overall?"
## [15] "How many years of professional work experience do you have in your field?"
## [16] "What is your highest level of education completed?"
## [17] "What is your gender?"
## [18] "What is your race? (Choose all that apply.)"
```

Per una millor visualització, modifiquem el nom de les columnes (features)

```
#{r canviem nom columnes, echo=FALSE}
nova_capçalera <- c("timestamp", "edat"
                    , "sector",
                    "lloc_feina","comentaris_lloc","salari_temps_complert",
                    "extres","moneda","altres_monedes","comentaris_extres",
                    "pais","US_estat","ciutat","experiencia_general_interval",
                    "experiencia_especifica_interval","nivell_estudis",
```

```
                    "genere","raca")

salaris_df <- rbind(nova_capçalera, salaris_df)
colnames(salaris_df) <- as.character(unlist(salaris_df[1, ]))
salaris_df <- salaris_df[-1, ]
```

```
colnames(salaris_df)
```

```
##  [1] "timestamp"                          "edat"
##  [3] "sector"                             "lloc_feina"
##  [5] "comentaris_lloc"                    "salari_temps_complert"
##  [7] "extres"                             "moneda"
##  [9] "altres_monedes"                     "comentaris_extres"
## [11] "pais"                               "US_estat"
## [13] "ciutat"                             "experiencia_general_interval"
## [15] "experiencia_especifica_interval" "nivell_estudis"
## [17] "genere"                             "raca"
```

Un cop carregat el dataset, farem una primera ullada a les dades per fer-nos una idea. Ho dividirem en 4 taules.

Taula 1

timestamp	edat	sector	lloc_feina
4/27/2021 11:02:10	25-34	Education (Higher Education)	Research and Instruction Librarian
4/27/2021 11:02:22	25-34	Computing or Tech	Change & Internal Communications Manager
4/27/2021 11:02:38	25-34	Accounting, Banking & Finance	Marketing Specialist
4/27/2021 11:02:41	25-34	Nonprofits	Program Manager
4/27/2021 11:02:42	25-34	Accounting, Banking & Finance	Accounting Manager
4/27/2021 11:02:46	25-34	Education (Higher Education)	Scholarly Publishing Librarian

Taula 2

comentaris_lloc	salari_temps_complert	extres	moneda	altres_monedes	comentaris_extres
NA	55000	0	USD	NA	NA
NA	54600	4000	GBP	NA	NA
NA	34000	NA	USD	NA	NA
NA	62000	3000	USD	NA	NA
NA	60000	7000	USD	NA	NA
NA	62000	NA	USD	NA	NA

Taula 3

pais	US_estat	ciutat
United States	Massachusetts	Boston
United Kingdom	NA	Cambridge
US	Tennessee	Chattanooga
USA	Wisconsin	Milwaukee
US	South Carolina	Greenville
USA	New Hampshire	Hanover

Taula 4

experiencia_general_interval	experiencia_especifica_interval	nivell_estudis	genere	raca
5-7 years	5-7 years	Master's degree	Woman	White
8 - 10 years	5-7 years	College degree	Non-binary	White
2 - 4 years	2 - 4 years	College degree	Woman	White
8 - 10 years	5-7 years	College degree	Woman	White
8 - 10 years	5-7 years	College degree	Woman	White
8 - 10 years	2 - 4 years	Master's degree	Man	White

I extreurem un petit resum de les columnes, valors únics, tipus de dades, NAs inicials.

	Num_NAs	Mida_Bytes	Valors_Unics	Tipus_Dada
timestamp	0	2253896	25359	character
edat	0	225584	7	character
sector	77	312792	1134	character
lloc_feina	0	1303608	13452	character
comentaris_lloc	20866	1011232	6963	character
salari_temps_complert	0	430912	3673	character
extres	7333	272952	852	character
moneda	0	225792	11	character
altres_monedes	27921	233616	122	character
comentaris_extres	25083	645672	2975	character
pais	0	246664	324	character
US_estat	5046	235576	137	character
ciutat	25	506544	4269	character
experiencia_general_interval	0	225704	8	character
experiencia_especifica_interval	0	225704	8	character
nivell_estudis	228	225640	6	character
genere	174	225568	5	character
raca	183	232696	51	character

En aquest punt, ja podem veure que el dataset conté molta informació que no és correcte degut a la introducció manual de dades que permet el formulari en moltes preguntes.

Pegunta 2.Integració, selecció de dades d'interés

1. Selecció dades d'interés
2. Mostrar resum variables.

RESPOSTA:

- Integració/fusió de múltiples datasets: No apliquem can fusió ni integració al tenir només una font de dades.
- Selecció de les dades:

El dataset és una enquesta orientada als Estats Units, però és possible que hi hagin altres països presents. Fem una petita exploració inicial, un cop hem passat tot a MAJÚSCULES i sense espais.

```
## # A tibble: 10 x 2
##    pais                    n
##    <chr>               <int>
##  1 UNITED STATES       10035
##  2 USA                  9073
##  3 US                   2782
```

```
##  4 CANADA                    1681
##  5 UK                         692
##  6 UNITED KINGDOM             634
##  7 U.S.                       606
##  8 UNITED STATES OF AMERICA   492
##  9 AUSTRALIA                  390
## 10 GERMANY                    197
```

Com podem veure, la majoria de registres es corresponen als Estats Units, així que farem l'estudi dels salaris dins d'aquest països. Per fer-ho, unificarem el nom del països i eliminarem la resta de registres i eliminarem la variable pais

També eliminarem la variable US_estat, ciutat, moneda (l'estudi serà pel Estats Units), i els camps opcionals de dades del formulari com comentaris_lloc,comentaris_extres,altres_monedes

Ara ens fixarem ara amb la variable sector, que l'agruparem per poder veure els sectors més representats:

```
## # A tibble: 10 x 2
##    sector                                n
##    <chr>                             <int>
##  1 COMPUTING OR TECH                  3730
##  2 NONPROFITS                         2121
##  3 EDUCATION (HIGHER EDUCATION)       2083
##  4 HEALTH CARE                        1627
##  5 ACCOUNTING, BANKING & FINANCE      1475
##  6 ENGINEERING OR MANUFACTURING       1428
##  7 GOVERNMENT AND PUBLIC ADMINISTRATION  1410
##  8 LAW                                 962
##  9 MARKETING, ADVERTISING & PR         919
## 10 EDUCATION (PRIMARY/SECONDARY)       717
```

Decidim quedar-nos amb els 8 més importants

Tornarem a visualitzar la taula de nou:

```
## # A tibble: 8 x 2
##   sector                                n
##   <chr>                             <int>
## 1 COMPUTING OR TECH                  3730
## 2 NONPROFITS                         2121
## 3 EDUCATION (HIGHER EDUCATION)       2083
## 4 HEALTH CARE                        1627
## 5 ACCOUNTING, BANKING & FINANCE      1475
## 6 ENGINEERING OR MANUFACTURING       1428
## 7 GOVERNMENT AND PUBLIC ADMINISTRATION  1410
## 8 LAW                                 962
```

PREGUNTA 3. Neteja de les dades

Avaluar duplicats: Mirem si hi han duplicats de dades:

```
duplicats <- salaris_df[duplicated(salaris_df),]
print (nrow(duplicats))
```

```
## [1] 0
```

Avaluar elements buits(NAs):

Ara busquem si hi han elements buits, dades amb zero valors o altres valors numèrics que indiquin la pèrdua de dades. Per fer-ho, visualitzarem una taula amb tots els NAs i el percentatge que significa dins de les dades

Evaluem els valors NAs del dataset.

	Columna	N_NA	Percentatge
extres	extres	3866	26.06
raca	raca	107	0.72
nivell_estudis	nivell_estudis	99	0.67
genere	genere	98	0.66
timestamp	timestamp	0	0.00
edat	edat	0	0.00

Imputació de valors:

De les dades de la taula, podem observar un nombre molt elevats de extres (26%) No es tracten de valors faltants en sí, sinó que l'enquestat no feia hores extres.

Imputarem el valor de zero a aquests als valors faltants.

Ara passarem a imputar els altres valors amb mètodes probabilístics. En aquest cas, farem servir KNN del paquet VIM.

```
kNN1.salaris_df<-kNN(salaris_df, k=3)
```

Veiem el resultat final

	Columna	N_NA	Percentatge
timestamp	timestamp	0	0
edat	edat	0	0
sector	sector	0	0
lloc_feina	lloc_feina	0	0
salari_temps_complert	salari_temps_complert	0	0
extres	extres	0	0

Eliminem les variables d'imputació (xxx_imp)
que son produïdes per la funció kNN1 (ens indica a quines columnes s'han hagut d'imputar valors).

Eliminem les variables d'imputació (xxx_imp) que son produïdes per la funció kNN1 (ens indica a quines columnes s'han hagut d'imputar valors)

```
kNN1.salaris_df <- kNN1.salaris_df[, !grepl("_imp$", names(kNN1.salaris_df))]
salaris_df<-kNN1.salaris_df[]
rm(kNN1.salaris_df)
```

Variable moneda:

Unifiquem la moneda a EUROS

```
# Factors de conversió a EUR

factor_conversio=0.92
#
# Conversió

# Creem la nova columna amb el salari convertit a euros
salaris_df <- salaris_df %>%
  mutate(salari_temps_complert_euros = as.numeric(salari_temps_complert) * factor_conversio)

salaris_df <- salaris_df %>%
  mutate(extres_euros = as.numeric(extres) * factor_conversio)
```

```
salaris_df <- subset(salaris_df, select = -salari_temps_complert)
salaris_df <- subset(salaris_df, select = -extres)
rm(factor_conversio)
```

- Modifiquem el format variable timestamp: No ens interessa ni el día ni la hora de l'enquesta, només ens quedarem amb el mes i l'any

```
salaris_df$timestamp <- as.POSIXct(salaris_df$timestamp, format = "%m/%d/%Y %H:%M:%S")
salaris_df$mes_any <- as.factor(format(salaris_df$timestamp, "%m/%Y"))
salaris_df <- subset(salaris_df, select = -timestamp)
```

- Passem a factors la resta de variables. Fixarem un màxim de 15 categries per variable

```
library(dplyr)
library(forcats)

salaris_df <- salaris_df %>%
  mutate(across(
    .cols = where(~ !is.numeric(.x)),
    .fns = ~ fct_lump(as.factor(.x), n = 15)
  ))
```

Valors extrems

Començarem l'estudi de possibles outliers.Per fer-ho, mirarem els quartils de les variables numèriques.

```
summary(salaris_df[, sapply(salaris_df, is.numeric)])
```

```
##  salari_temps_complert_euros  extres_euros
##  Min.   :      0              Min.   :      0
##  1st Qu.:  55200              1st Qu.:      0
##  Median :  76820              Median :      0
##  Mean   :  89734              Mean   :  10008
##  3rd Qu.: 110400              3rd Qu.:   5520
##  Max.   :2392000              Max.   :1380000
```

Podem veure que:

Salari a temps complet (`salari_temps_complert_euros`)

- **Minim:** 0 Hi ha casos amb salari 0 (pot ser errors o persones sense salari base). Assumirem que l'enquesta tothom té salari, així que els considerarem com errors
- **1r quartil (Q1):** 51.713 €. Significa que el **25% més baix** dels registres guanya **menys de 51.713 €**.
- **Mediana (Q2):** 72.023 € La meitat dels registres tenen un salari **inferior a 72.023 €** i l'altra meitat superior.
- **Mitjana:** 85.176 € El salari **promig** és 85.176 €, superior a la mediana. Això ens indica que hi han valors alts que estàn afectant a la mitjana (possibles outliers).
- **3r quartil (Q3):** 103.040 € . El **25% més alt** guanya **més de 103.040 €**.
- **Màxim:** 9.200.000 €. Hi ha un salari extremadament alt, que podria ser una dada errònia o corresponent a un alt directiu.

Haurem de limitar el sou màxim per no distorsionar l'estudi.

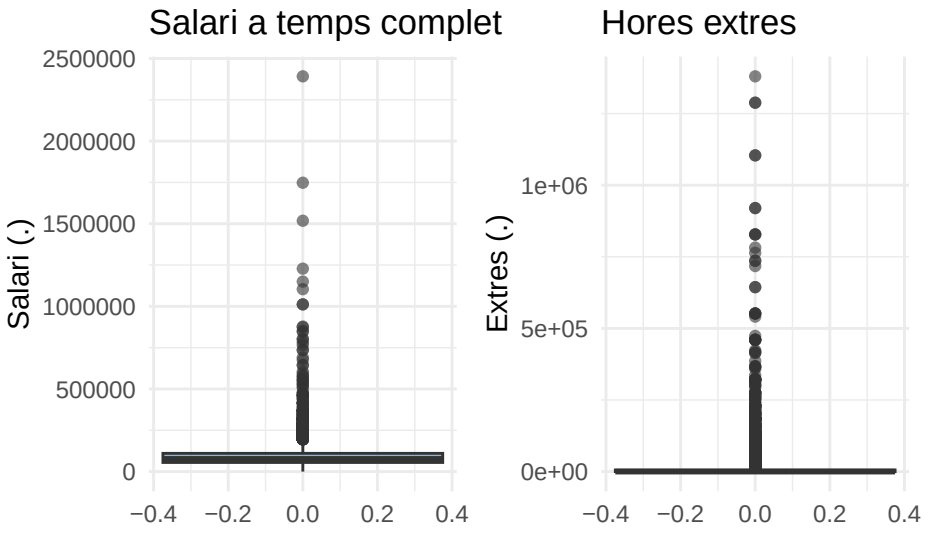
Extres (`extres_euros`)

- **Minim:** 0 Hi ha persones que no tenen extres (el considerem normal)

- **1r quartil (Q1):** 0. El **25% més baix** no té cap extra.
- **Mediana (Q2):** 0 .La meitat dels registres no reben extres. Això vol dir que més de la meitat dels enquestats no reben hores extres
- **Mitjana:** 8.727 € . Tot i que la mediana és **0**, la mitjana és **8.727 €**, indicant que algunes persones tenen extres alts que pugen el promig (seria un cas similar a l'anterior del salari). Dades no normalment distribuïdes.
- **3r quartil (Q3):** 5.520 € → El **25% superior** cobra més de 5.520 € en extres.
- **Màxim:** 1.380.000 € → Existeix un registre massa alt (o un cas especial o una dada incorrecta).

La distribució dels ingressos extres sembla **molt asimètrica**, amb una **majoria de registres a 0**, però amb alguns casos amb **valors molt alts** que eleven la mitjana. Analitzarem els possibles outliers que puguin distorsionar l'anàlisi.

Per fer-ho, farem una exploració visual;



Com podem veure, els boxplots no apareixen correctament dibuixats. Això és degut a l'existència de valors massa grans. Farem una coversió logarítmca de les dades

Variable salari:

La sortida mostra un valor màxim extremadament alt (4.080e+09), que ens indica la presència d'outliers, afectant a la mitjana que és molt més alta que la mediana.

Visualment, podem veure com la representació ens demostra presència outliers.

Per tot això, el nostre estudi es centrarà en sous que es considerin "més normals", valors del rang de [15.000,150.000] euros.

```
print(paste("Registres < 15.000 euros:", num_regs_baix_salari))
```

```
## [1] "Registres < 15.000 euros: 70"
```

```
print(paste("Registres > 150.000 euros:", num_regs_alt_salari))
```

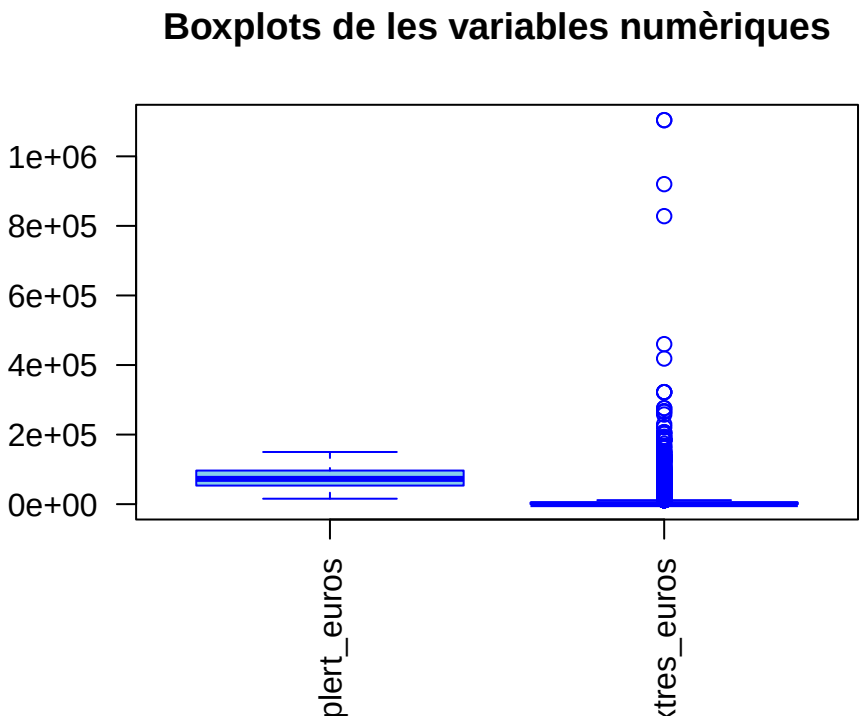
```
## [1] "Registres > 150.000 euros: 1492"
```

```
salaris_df_filtrat <- salaris_df %>%
  filter(salari_temps_complert_euros >= 15000 & salari_temps_complert_euros <= 150000)
```

Tornem a avaluar els resultats:

```
##  salari_temps_complert_euros  extres_euros
##  Min.   : 15640              Min.   :      0
```

```
##  1st Qu.: 53360             1st Qu.:      0
##  Median : 72680             Median :      0
##  Mean   : 76927             Mean   :   5304
##  3rd Qu.: 96600             3rd Qu.:   4600
##  Max.   :149960             Max.   :1104000
```



Eliminems el outliers

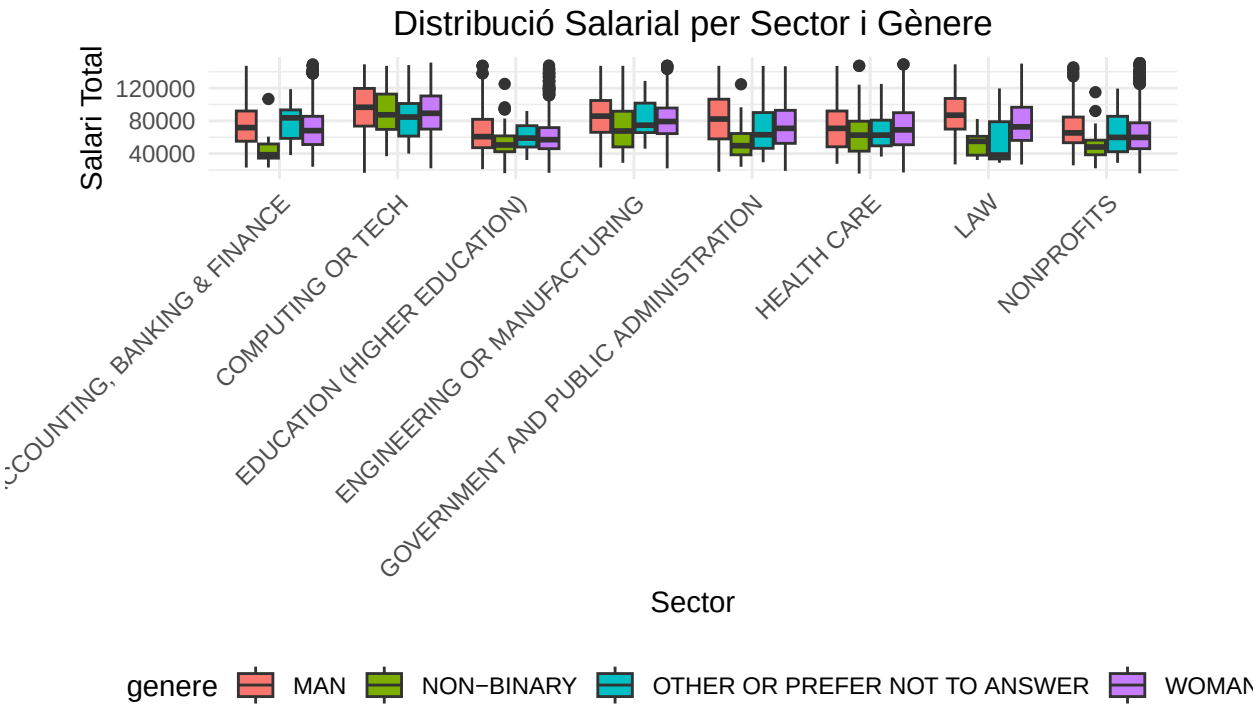
```
##  salari_temps_complert_euros  extres_euros
##  Min.   : 15640               Min.   :    0
##  1st Qu.: 51520               1st Qu.:    0
##  Median : 69000               Median :    0
##  Mean   : 72761               Mean   : 1505
##  3rd Qu.: 90160               3rd Qu.: 1840
##  Max.   :149960               Max.   :11500
```

Finalment, construim una nova variable amb el salari total. Ens servirà com a variable objectiu en posteriors anàlisis.

També farem un altre columna amb el logaritmr del salari. Aquesta variable la farem servir en els models d'aprenentatge posteriors perque, donat la disparitat de salaris, és preferible entrenar els models fent servir la versió logarítmica del sou.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15640   52440   69662   74082   92000  151156
```

Finalment, i abans de començar l'anàl · lis amb models, treurem algunes gràfiques de salari per sector i gènere.



Visualment podem veure com, independentment del sector, els homes guanyen més que les dones.

PREGUNTA 4: MODEL SUPERVISAT I NO SUPERVISAT.

MODEL SUPERVISAT: RANDOM FOREST

Farem servir RandomForest que és una versió millorada dels decissions trees (son models combinats a base de decission trees).

Solen funcionar millor que els model de regressió lineal i les variables presenten relacions "amagades" que la combinació lineal de la regressió no captura.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x purrr::%||%()    masks base::%||%()
## x purrr::compose() masks flextable::compose()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(randomForest)
```

```
## randomForest 4.7-1.1
## Type rfNews() to see new features/changes/bug fixes.
##
## Adjuntando el paquete: 'randomForest'
##
## The following object is masked from 'package:ggplot2':
##
##     margin
##
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
salaris_df_sup <- salaris_df[]

# Filtrar i preparar el dataset
salaris_df_sup <- salaris_df_sup %>% drop_na(sou_total_log)  # Eliminem files amb valors perduts

salaris_df_sup <- salaris_df_sup %>%
  mutate(across(c(edat, sector, lloc_feina, nivell_estudis, genere, raca), as.factor))

set.seed(123) # Per fer el resultat replicable

# Divisió en train/test (80% entrenament, 20% test)
train_index <- sample(seq_len(nrow(salaris_df_sup)), size = 0.8 * nrow(salaris_df_sup))
train <- salaris_df_sup[train_index, ]
test <- salaris_df_sup[-train_index, ]

model_rf <- randomForest(sou_total_log ~ ., data = train, ntree = 500, mtry = 3, importance = TRUE)

# Veure importància de les variables
importance(model_rf)
```

```
##                                   %IncMSE IncNodePurity
## edat                            18.511397     11.770061
## sector                          23.409979     86.420230
## lloc_feina                      17.865169     25.018963
## experiencia_general_interval    17.626995     16.295146
## experiencia_especifica_interval 20.402381     47.506518
## nivell_estudis                  18.295958     26.020042
## genere                           8.327919      7.396156
## raca                             4.676274      8.571641
## mes_any                          2.748625      6.288031
## sou_total                      304.216841   1246.305830
```

```r
pred_test <- predict(model_rf, newdata = test)

# Avaluar el rendiment amb MSE i R²
mse_rf <- mean((pred_test - test$sou_total_log)^2)
r2_rf <- 1 - (sum((pred_test - test$sou_total_log)^2) / sum((test$sou_total_log - mean(test$sou_total_l

cat("MSE:", mse_rf, "\nR²:", r2_rf)
```

```
## MSE: 0.0007780665
## R²: 0.995103
```

El R2 = 0.9953 indica que el model explica gairebé tota la variabilitat del salari, i el MSE = 0.00075 és extremadament baix, indicant que l'error en les prediccions és molt petit. Per entendre la sortida del model, hem de recordar que %IncMSE l'mpacte en l'error de predicció (com més alt, més important és la variable).

Per tant, les variables més influents en el salari són el sector, l'edat (com més edat millor sours), experiència, rol.

MODEL NO SUPERVISAT:

Farem servir el mètode de clustering, que és un dels més utilitzats i en el que les dades s'agrupen dins d'un mateix clúster quan tenen una alta similitud i alhora són fàcilment distingibles d'altres grups. Farem servir FAMD, ja que tenim totes excepte una variables categòriques. L'anàlisi es fa sobre sobre el dataset 'salaris_df_no_sup'.

Per problemes de rendiment en l'execució s'ha reduït el dataset a 1000 registres.

```
if(!require(cluster)){
  install.packages("cluster")
}
```

```
## Cargando paquete requerido: cluster
```

```
if(!require(factoextra)){
  install.packages("factoextra")
}
```

```
## Cargando paquete requerido: factoextra
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
if(!require(dplyr)){
  install.packages("dplyr")
}
if(!require(uwot)){
  install.packages( "uwot")
}
```

```
## Cargando paquete requerido: uwot
```

```
## Cargando paquete requerido: Matrix
```

```
##
## Adjuntando el paquete: 'Matrix'
```

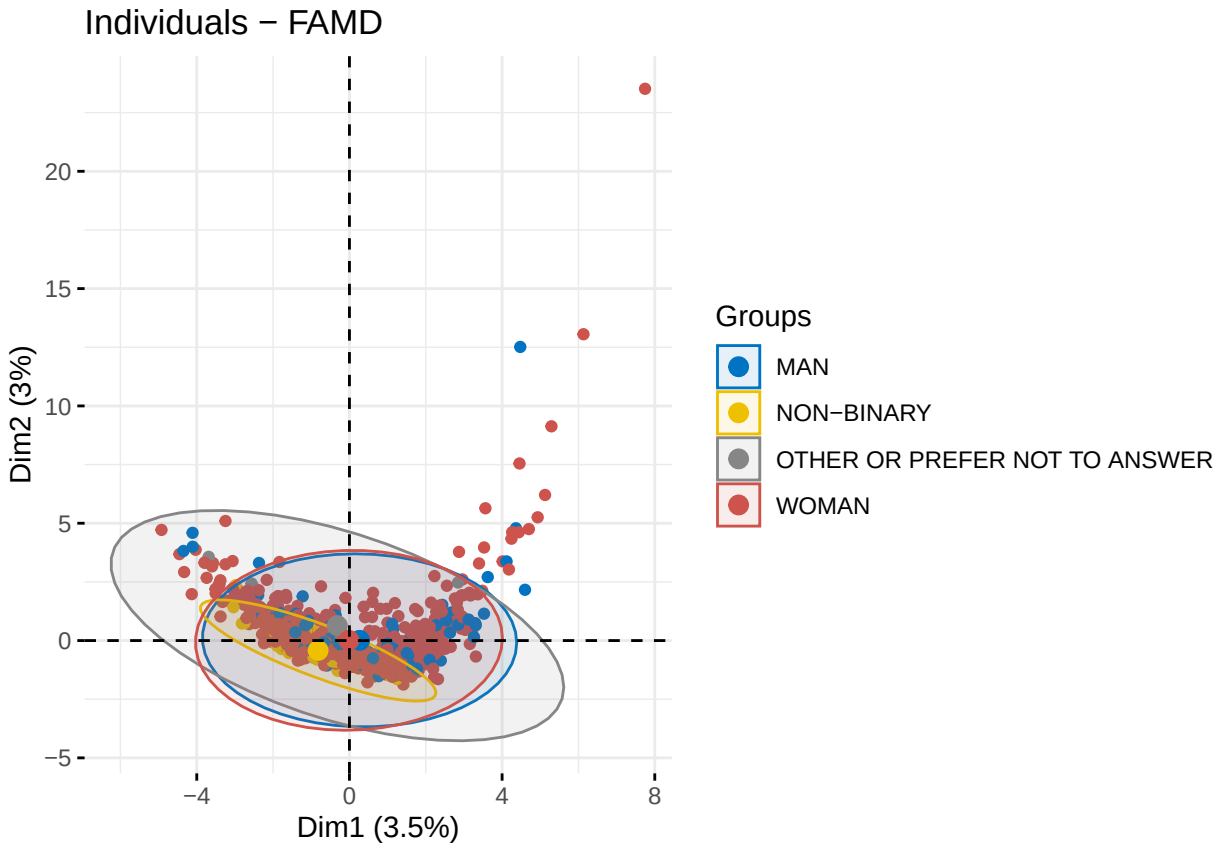
```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
if(!require(FactoMineR)){
  install.packages( "FactoMineR")
}
```

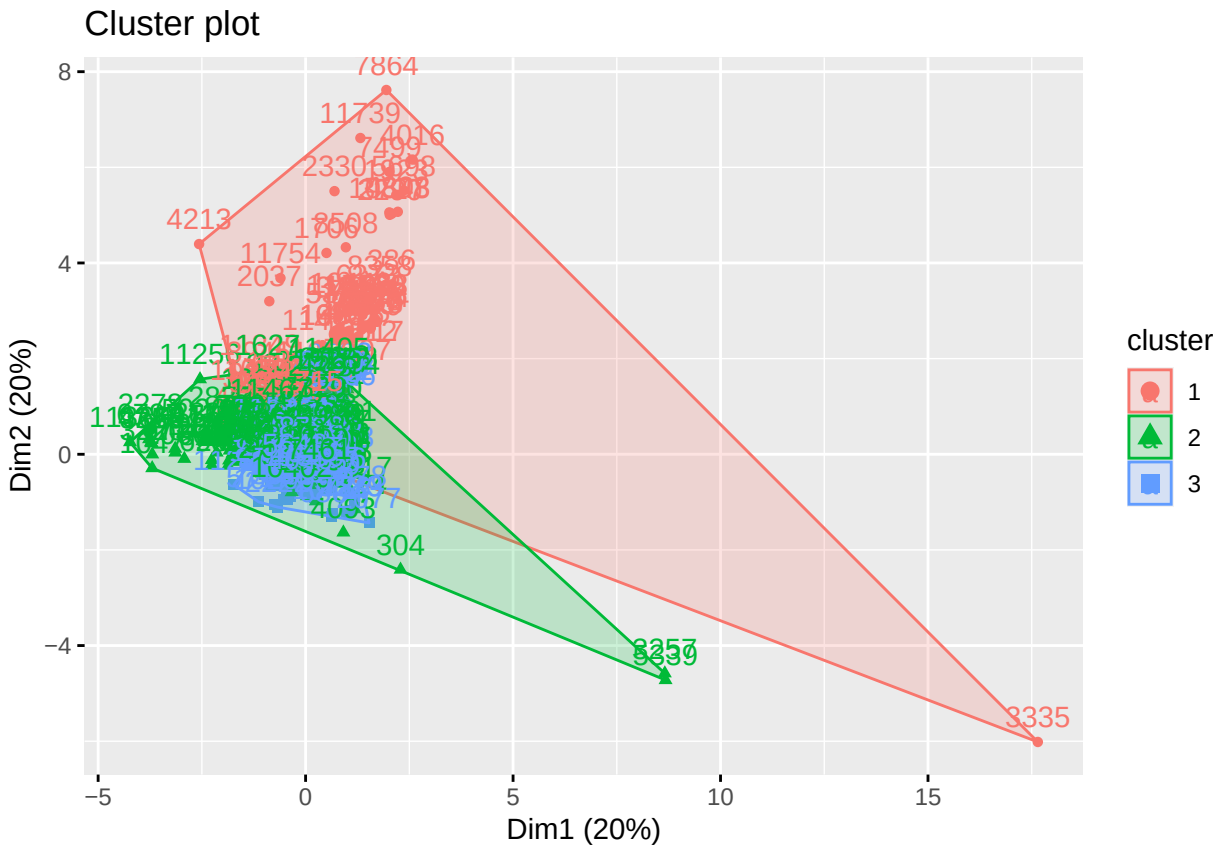
```
## Cargando paquete requerido: FactoMineR
```

```
# Cargar las librerías
library(cluster)
library(factoextra)
library(dplyr)
library(uwot)
library(FactoMineR)

# FArm un subset aleatori de 1000 registres perquè hi ha problemes en l'execució
set.seed(123)  # Per fer la mostra reproductible
df_sub <- salaris_df_no_sup[sample(1:nrow(salaris_df_no_sup), 1000), ]
df_sub$sou_total <- cut(df_sub$sou_total, breaks = 3, labels = c("baix", "mitjà", "alt"))

# Fer MCA sobre variables categòriques
mca_res <- MCA(df_sub %>% select(-sou_total), graph = FALSE)
```

```r
res_famd <- FAMD(df_sub, graph = FALSE)
#  Visualitzar individus
fviz_mca_ind(res_famd,
             label = "none",
             habillage = df_sub$genere,
             addEllipses = TRUE,
             palette = "jco")
```

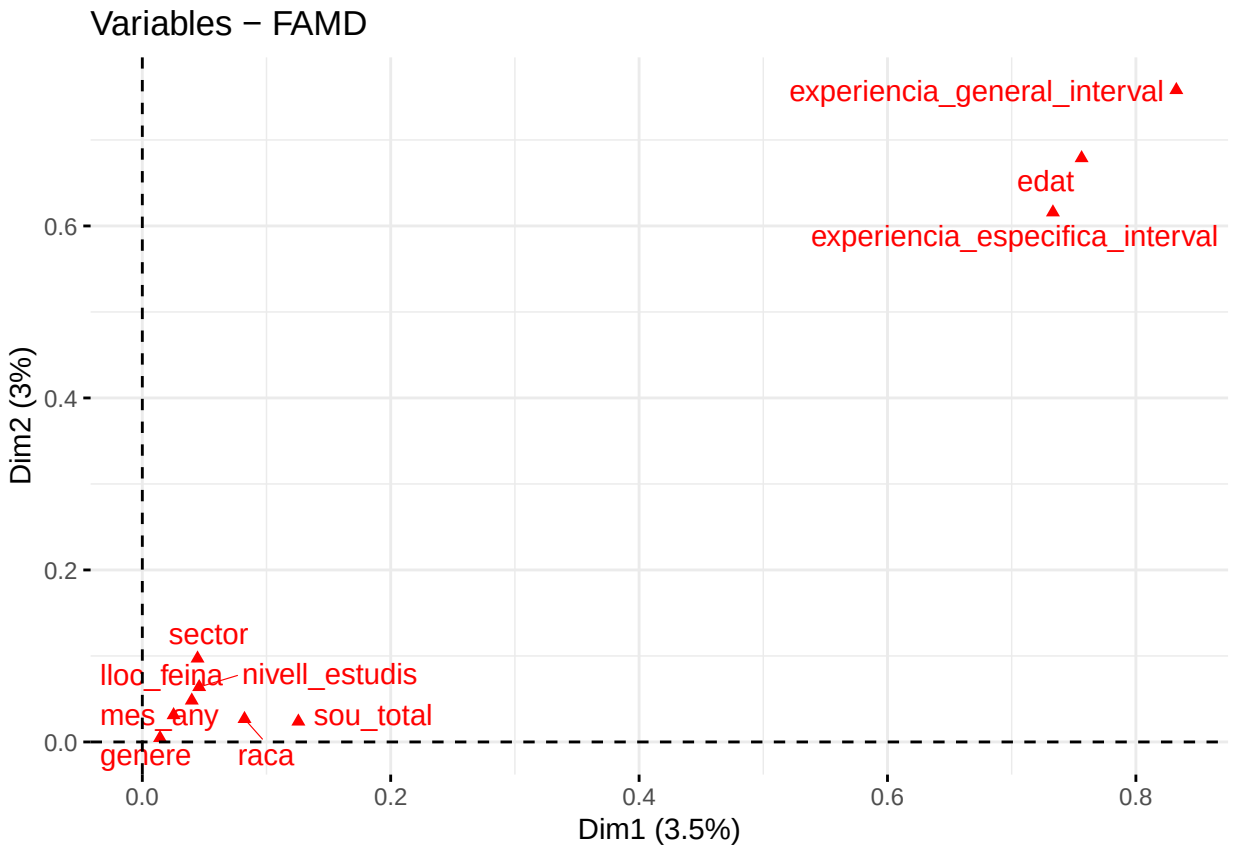



```r
# Clustering sobre components MCA (exemple amb k-means)
mca_vars <- res_famd$ind$coord  # Coordenades individuals

k_res <- kmeans(mca_vars, centers = 3)
fviz_cluster(list(data = mca_vars, cluster = k_res$cluster))
```



Mirem com contribueixen les variables:

```
fviz_famd_var(res_famd, repel = TRUE)
```



Aplica una prova per contrast d'hipòtesi

Plantegem la següent pregunta: "La mitjana del salari total anual a temps complert dels homes de raca blanca són més grans que la de la resta dels grups de persones".

Hipòtesi del contrast

La hipòtesi nul · la i alternativa serien:

H0 : u1 - u2 >= 0 (el salari mitjà dels homes blancs no és més gran: hipòtesi nul · la) H1 : u1 - u2 < 0 (el salari mitjà dels homes blancs és més gran: hipòtesi alternativa) on u1 denota la mitjana del salari dels homes de raça blanca i u2 la mitjana del salari de la resta de grups de persones.

Test a aplicar: Pel teorema del límit central, podem assumir normalitat, ja que tenim una mostra amb moltes observacions i es vol fer un test sobre la mitjana. Per tant, apliquem un test d'hipòtesis de dues mostres sobre la mitjana. Aplicarem la distribució t, atès que no es coneix la variància de la població. Es tracta de dues mostres independents amb variància desconeguda i diferent. És un test unilateral. Considerarem un nivell de confiança del 99%.

Crearem les dues mostres, una per a cada conjunt de població.

```
df_1 <- subset(salaris_df_hipotesis, genere == "MAN" & raca == "WHITE" & sou_total < 150000 & sou_total
df_2 <- subset(salaris_df_hipotesis, (!(genere == "MAN" & raca == "WHITE" ) & sou_total < 150000 & sou_
```

Apliquem el test:

```
t.test(df_1, df_2, alternative="greater", var.equal=FALSE, conf.level=0.99)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  df_1 and df_2
## t = 15.377, df = 1973, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 99 percent confidence interval:
##  10736.39      Inf
## sample estimates:
## mean of x mean of y
##  85038.08  72386.10
```

El p-value és molt inferior a 0.01, (ja que usem un 99% de confiança), rebutgem clarament la hipòtesi nul · la. L'interval de confiança (el valor que està a Inf) és positiu i la diferència és significativa. La mitjana x (homes de raça blanca) és aproximadament 8.000 euros més alt que la mitjana y.

Resposta al test d'hipòtesi:

El salari mig dels homes de raça blanca és superior al de la resta de grups poblacionals.