PAC2

Carlos A. García

November 11, 2019

Títol de la visualització on és presenten el dataset o datasets escollits

Diferències salarials per sexe i per lloc de feina

Descripció curta del document i del que s'hi presenta

Les dades mostren les diferències salarials entre homes i dones per a un mateix lloc de feina i categoria laboral. Les dades estan detallades per país (Estats Units i Regne Unit) i agrupades per categoria laboral.

Els valors estan especificats en la moneda local (Dólars per a les dades dels EUA i Lliures Esterlines per a les del RU); una petita dificultat afegida és convertir a una única moneda; en el nostre cas, Euros. Les dades són de 2014 i estan extretes del "Bureau of Labor Statistics[1]" (Estats Units) i de la "Office for National Statistics[2]" (Regne Unit).

Les dades són per a empleats a temps complet; no s'inclouen ni els treballadors a temps parcial ni els freelance. Els valors monetaris es corresponen amb mitges anuals.

Les dades[3] originals es poden trobar a la web[4].

Les dades, presentació: Què en sabeu de les dades: tipus, estructura, curiositats

Les dades originals són:

- **Occupation.** Lloc de feina. Dada alfanumèrica.
- **Category.** Categoria del lloc de feina. Funciona com a aglutinador. Dada alfanumèrica.
- **Women average annual salary ($).** Salari anual mitjá de les dones per al lloc de feina especificat. Expressat en la moneda del país. Variable numèrica.
- **Men average annual salary ($).** Salari anual mitjá dels homes per al lloc de feina especificat. Expressat en la moneda del país. Variable numèrica.
- **Pay gap ($).** Diferència entre el salari dels homes i de les dones. Un valor positiu indica que els homes guanyen més. Negatiu, que són les dones qui més guanyen. Variable numèrica.
- **Pay gap as a percentage.** Diferència de salari expresada en percentatge. Variable numèrica.

A més, hi ha ha dues variables implícites que hem incorporat al dataset:

- **País.** País de la mostra. Dada alfanumèrica. Pot ser Estats Units o Regne Unit.
- **Moneda.** Moneda de la mostra. Dada alfanumèrica. Pot ser Dólar o Lliura Esterlina.

[1] https://www.bls.gov/
[2] https://www.ons.gov.uk/
[3] https://docs.google.com/spreadsheets/d/1Qih5qBcuTntLbx7G7BzunRSOgGD0b_zc07sTzqiKGn4/edit#gid=1275614270
[4] https://informationisbeautiful.net/visualizations/gender-pay-gap/

Les dades, exploració. Què hi heu descobert: evidències, tendencies, outlayers

Les dades, procediment i eines. Explicar com ho heu descobert: amb quines eines, amb quines operacions

Lectura i tractament inicial de les dades

Carreguem les dades del dataset original (incorporant les variables de país i moneda)

```
salaryGap <- read.csv2("salaryGap.csv", header = TRUE, sep = ",", dec = ".")
```

Canviem el nom de les columnes a un més adient. Les originals incorporen símbols extranys.

```
names(salaryGap)[names(salaryGap) == "ï..Occupation"] <- "Occupation"
names(salaryGap)[names(salaryGap) == "Women.average.annual.salary...."] <- "WomenAverageAnnualSalary"
names(salaryGap)[names(salaryGap) == "Men.average.annual.salary...."] <- "MenAverageAnnualSalary"
names(salaryGap)[names(salaryGap) == "Pay.gap...."] <- "PayGap"
names(salaryGap)[names(salaryGap) == "Pay.gap.as.a.percentage"] <- "PayGapAsAPercentage"
salaryGap["WomenAverageAnnualSalaryEUR"] <- salaryGap["WomenAverageAnnualSalary"]
```

Calculem les columnes en EUR (no és possible comparar diferentes monedes)

```
chageUSDEUR = 0.91
chageUKPEUR = 1.17
salaryGap$WomenAverageAnnualSalaryEUR[salaryGap$Currency == "USD"] <-
  (salaryGap$WomenAverageAnnualSalary) * chageUSDEUR
salaryGap$WomenAverageAnnualSalaryEUR[salaryGap$Currency == "UKP"] <-
  (salaryGap$WomenAverageAnnualSalary) * chageUKPEUR
salaryGap$MenAverageAnnualSalaryEUR[salaryGap$Currency == "USD"] <-
  (salaryGap$MenAverageAnnualSalary) * chageUSDEUR
salaryGap$MenAverageAnnualSalaryEUR[salaryGap$Currency == "UKP"] <-
  (salaryGap$MenAverageAnnualSalary) * chageUKPEUR
salaryGap$salaryGapEUR <- salaryGap$MenAverageAnnualSalaryEUR - salaryGap$WomenAverageAnnualSalaryEUR
salaryGapUS <- filter(salaryGap, Country == "US")
salaryGapUK <- filter(salaryGap, Country == "UK")
```

Validem que no hi ha nulls

```
sum(is.na(salaryGap$WomenAverageAnnualSalaryEUR))
```

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

```
sum(is.na(salaryGap$MenAverageAnnualSalaryEUR))
```

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

```
sum(is.na(salaryGap$salaryGapEUR))
```

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

```
sum(is.na(salaryGap$PayGapAsAPercentage))
```

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

Resum de les dades

Mostrem els resums de les variables numèriques en EUR

```
summary(salaryGap$WomenAverageAnnualSalaryEUR)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##    2692   24776   35159   38977   50710  115718
```

```
summary(salaryGap$MenAverageAnnualSalaryEUR)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3766   30069   43136   48177   62300  136647
```

```
summary(salaryGap$salaryGapEUR)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -2539    4184    7605    9200   12370   46482
```

```
summary(salaryGap$PayGapAsAPercentage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.1153  0.1520  0.2257  0.2524  0.3392  0.8716
```

Mostrem els resums de les variables alfanumèriques

```
summary(salaryGap$Currency)
```

```
## UKP USD
## 237 142
```

```
summary(salaryGap$Country)
```

```
##  UK  US
## 237 142
```

```
head(summary(salaryGap$Occupation), 10)
```

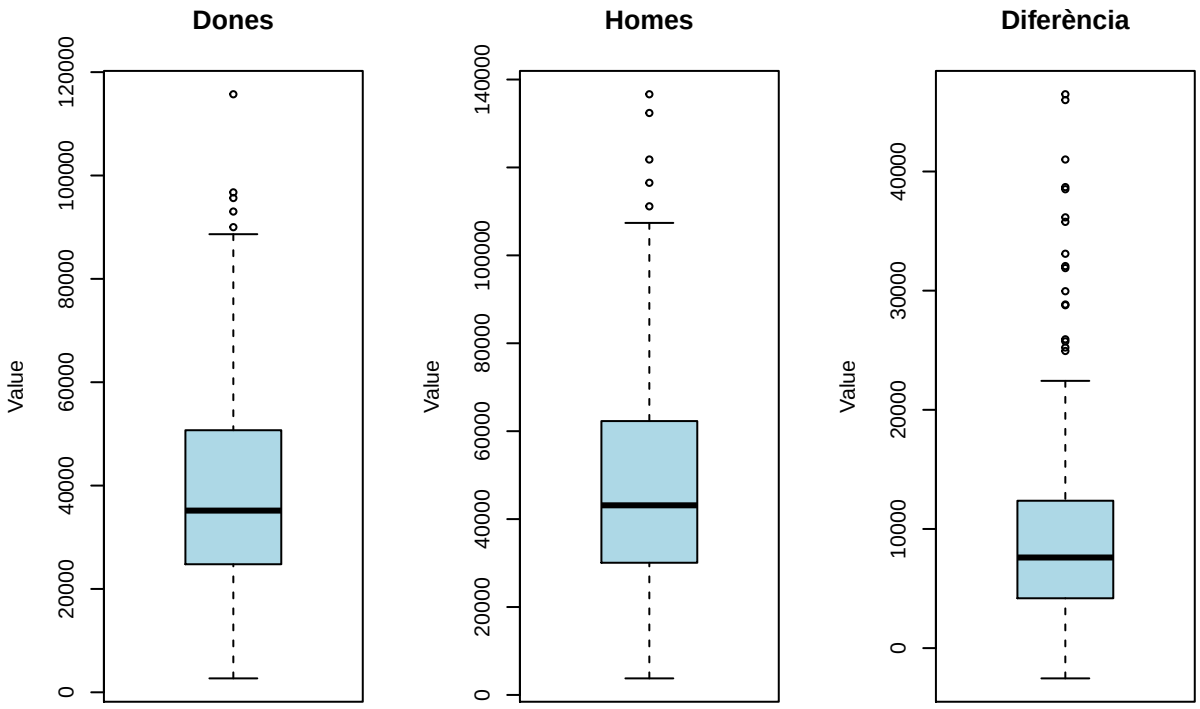
```
##               Admin        Construction Protective services
##                   3                   3                   3
##         Accountants        Arts & media              Bakers
##                   2                   2                   2
##    Care & education            Cashiers     Civil engineers
##                   2                   2                   2
##               Cooks
##                   2
```

```
summary(salaryGap$Category)
```

```
##        admin & organisation            care & education
##                          29                          48
##           creative & media               law & justice
##                          15                          20
##                manual work      sales & serving others
##                          40                         102
## science, tech & engineering     senior managers & execs
##                          72                          53
```

Gràficament, als boxplots, es veu clarament que els homes cobren més que les dones

```
par(mfrow=c(1,3))
boxplot(salaryGap$WomenAverageAnnualSalaryEUR,
                                        main="Dones", xlab="", ylab="Value", col="#ADD8E6")
boxplot(salaryGap$MenAverageAnnualSalaryEUR,
                                       main="Homes", xlab="", ylab="Value", col="#ADD8E6")
boxplot(salaryGap$salaryGapEUR,
                           main="Diferència", xlab="", ylab="Value", col="#ADD8E6")
```



Mostrem els outliers:

```
boxplot.stats(salaryGap$WomenAverageAnnualSalaryEUR)$out
```

```
## [1]  90002.64  96735.60 115717.68  93024.36  95640.48
```

```
boxplot.stats(salaryGap$MenAverageAnnualSalaryEUR)$out
```

```
## [1] 111154.7 121801.7 116508.6 132387.8 136646.6
```

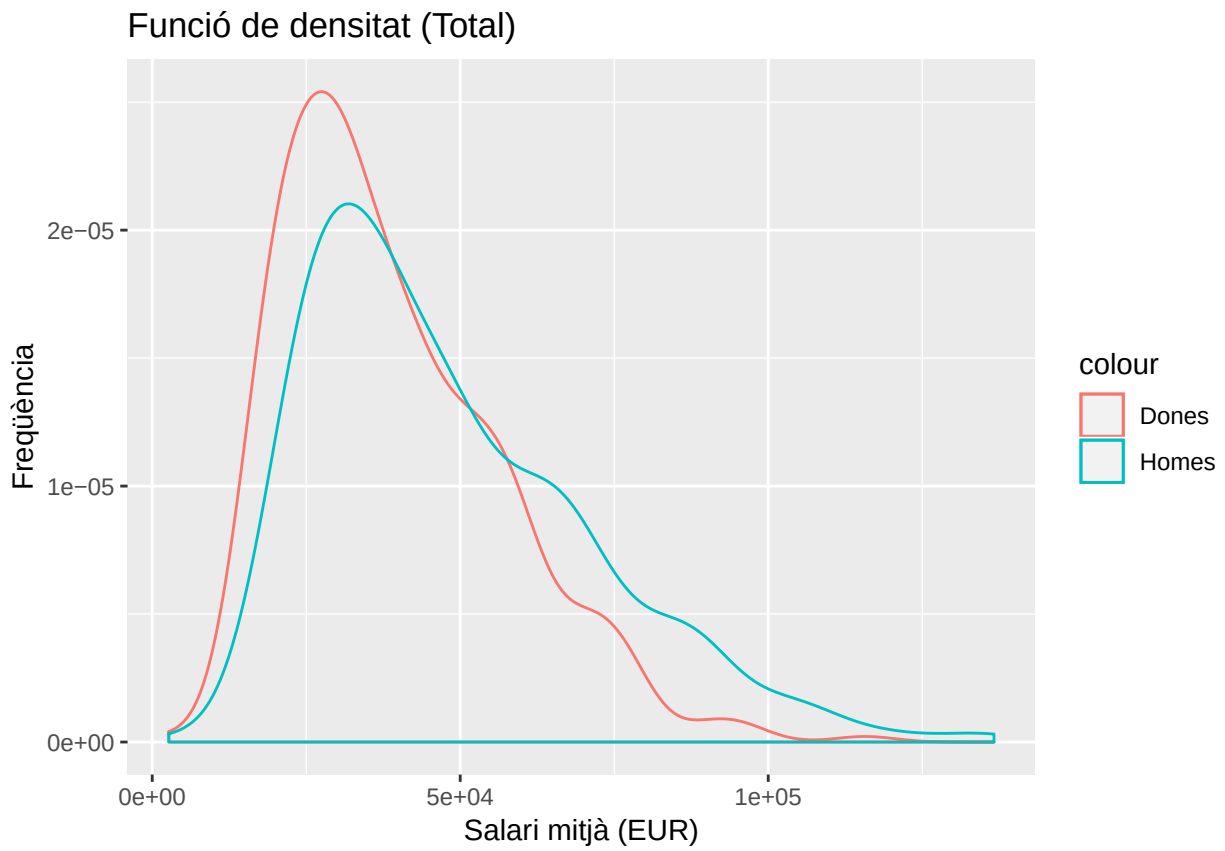
```
boxplot.stats(salaryGap$salaryGapEUR)$out
```

```
##  [1] 24937.64 35773.92 36152.48 29953.56 25742.08 31893.68 32062.68
##  [8] 45995.04 46481.76 28838.16 25229.88 38511.72 28777.32 33096.96
## [15] 41006.16 25917.84 38691.90
```

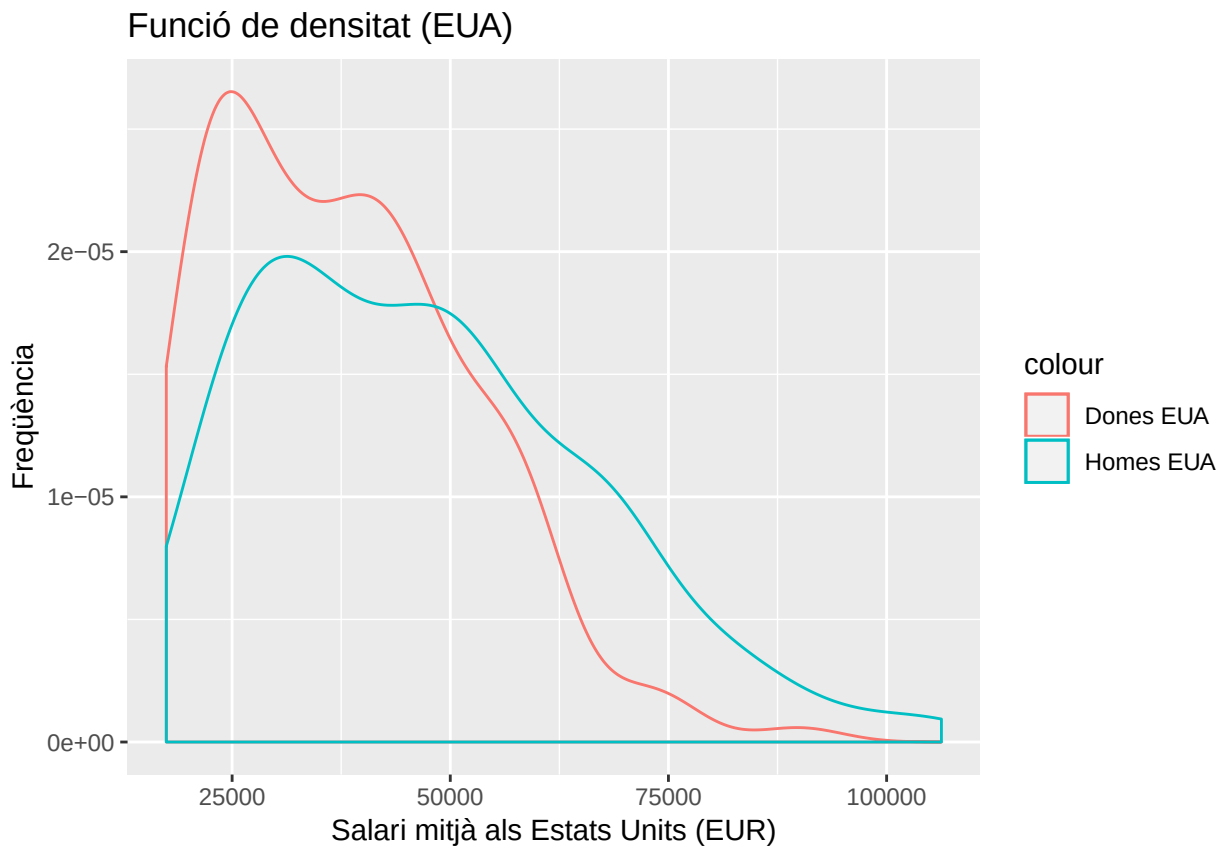
A la gràfica de densitat podem comprovar com les dades dels homes estan desplaçades a la dreta; això vol dir que trobem més valors masculins en el rang de salaris alts. També es pot veure com el pic de valors femenins és més alt i més a l'esquerra; això vol dir que hi ha moltes dones que guanyen poc.

Tot i així, es pot veure clarament que la diferència canvia depenent del país, tot i que es manté que els homes guanyen més.

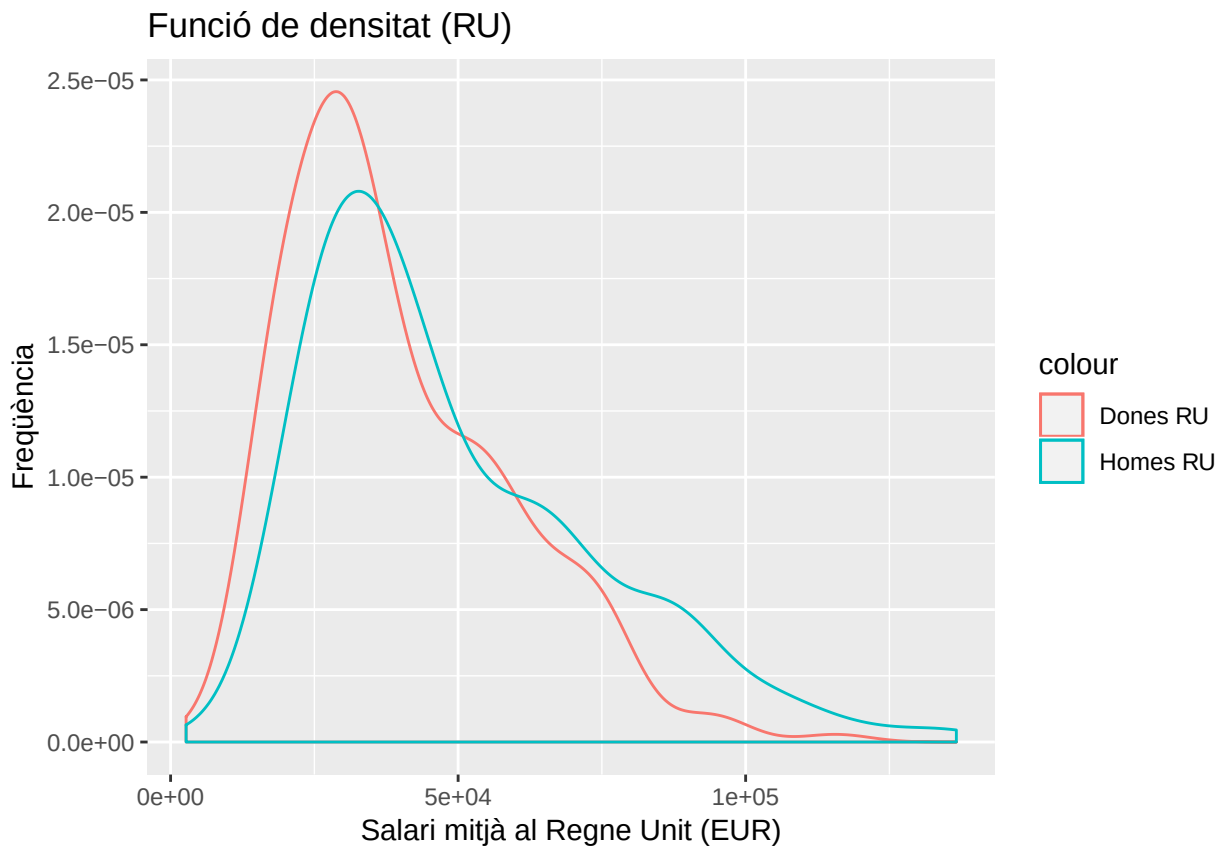
```
ggplot(salaryGap) +
  xlab("Salari mitjà (EUR)") + ylab("Freqüència") + ggtitle("Funció de densitat (Total)") +
  geom_density(aes(x = WomenAverageAnnualSalaryEUR, color = "Dones")) +
  geom_density(aes(x = MenAverageAnnualSalaryEUR, color = "Homes"))
```



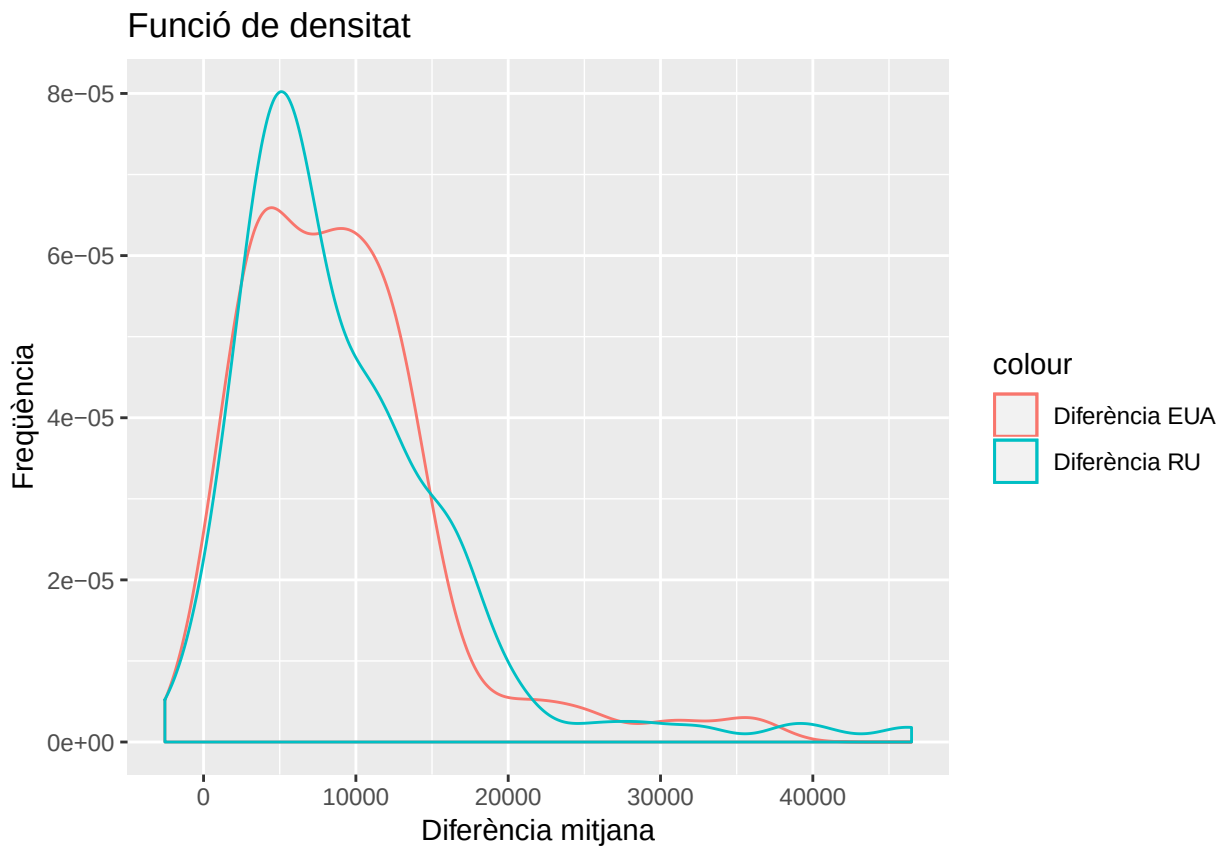

```
ggplot(salaryGapUS) +
  xlab("Salari mitjà als Estats Units (EUR)") + ylab("Freqüència") + ggtitle("Funció de densitat (EUA)")
  geom_density(aes(x = WomenAverageAnnualSalaryEUR, color = "Dones EUA")) +
  geom_density(aes(x = MenAverageAnnualSalaryEUR, color = "Homes EUA"))
```




```
ggplot(salaryGapUK) +
  xlab("Salari mitjà al Regne Unit (EUR)") + ylab("Freqüència") + ggtitle("Funció de densitat (RU)") +
  geom_density(aes(x = WomenAverageAnnualSalaryEUR, color = "Dones RU")) +
  geom_density(aes(x = MenAverageAnnualSalaryEUR, color = "Homes RU"))
```




```
par(mfrow=c(1,2))
ggplot() +
  xlab("Diferència mitjana") + ylab("Freqüència") + ggtitle("Funció de densitat") +
  geom_density(data=salaryGapUS, aes(x = salaryGapEUR, color = "Diferència EUA")) +
  geom_density(data=salaryGapUK, aes(x = salaryGapEUR, color = "Diferència RU"))
```



Com a curiositat, mostrem els valors (lloc de feina, categoria, paí, moneda, ...) pels quals una dona guanya el mateix o més de mitja que un home; només hi ha 6 casos:

```
salaryGapWM <- filter(salaryGap, WomenAverageAnnualSalaryEUR >= MenAverageAnnualSalaryEUR)
salaryGapWM <- select (salaryGapWM,
                       Occupation,  Category, Country,  Currency,
                       WomenAverageAnnualSalaryEUR, MenAverageAnnualSalaryEUR)
kable(salaryGapWM, caption = "Feines on les dones gunayen igual o més que els homes"
      ,col.names = c("Ocupació","Categoria","País", "Moneda", "Salari dones","Salari homes")
      ,align = c('r', 'r', 'r', 'r', 'r', 'r'))
```

Table 1: Feines on les dones gunayen igual o més que els homes

Ocupació	Categoria	País	Moneda	Salari dones	Salari homes
Stock clerks	sales & serving others	US	USD	24322.48	23849.28
Health technicians	science, tech & engineering	US	USD	29243.76	29243.76
Vocational trainers	sales & serving others	UK	UKP	31271.76	30663.36
Lorry drivers	sales & serving others	UK	UKP	37599.12	37599.12
Finance & business	senior managers & execs	UK	UKP	70049.07	67510.17
Finance technicians	senior managers & execs	UK	UKP	23426.91	21797.10

print(out, echo=FALSE)

Visualització sobre les dades. Un Dashboard o un conjunt de visualitzacions sobre els datasets escollits