PAC2

Carlos A. García

November 17, 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 6 variables i 379 files (237 registres del Regne Unit i 142 dels Estats Units):

- **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 evidències que hem trobat són:

- **Els homes guanyen més que les dones a la gran majoria dels llocs de feina (96.5%).** Els casos en els que una dona guanya igual o més de mitja són molt excepcionals (3.5
- **La diferència és major als Estats Units que al Regne Unit.** La gran majoria de salaris de dones als EUA es situa a la franja esquerra del rang total. Això vol dir que cobren molt menys. Les gràfiques del RU mostren que la gràfica dels homes està desplaçada a la dreta en comparació amb la de les dones. Això vol dir que, en general, cobren més.
- **Especialment dolorós és que a feines ocupades majoritàriament per dones amb un baix salari,** com pot ser "Cleaning occupations" trobem diferències del 87
- **Trobem diferències importants a tots els nivells salarials.** Que una dona arribi a directiu d'una empresa no vol dir que acabi guanyant el mateix que si for un home. Per exemple, un "Bank manager" masculí guanya de mitja més de 30.000€ més a l'any que una dona.

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

Hem obtingut les dades de "Information is Beautiful[5]" Com a eines hem utilitzat:

- **R i RStudio** per a l'anàlisi de dades.
- **LaTex** per a la realització del document.
- **RMarkdown** per a donar format al document.
- **Visme**[6] com a eina web de visualització de dades.

Les operacions realitzades són les bàsiques de qualsevol anàlisi estadístic. Les operacions concretes es detallen a continuació.

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"] <-
```

[5] https://informationisbeautiful.net/visualizations/gender-pay-gap/
[6] https://www.visme.co/

```
  salaryGap$WomenAverageAnnualSalary[salaryGap$Currency == "USD"] * chageUSDEUR

salaryGap$WomenAverageAnnualSalaryEUR[salaryGap$Currency == "UKP"] <-
  salaryGap$WomenAverageAnnualSalary[salaryGap$Currency == "UKP"] * chageUKPEUR

salaryGap$MenAverageAnnualSalaryEUR[salaryGap$Currency == "USD"] <-
  salaryGap$MenAverageAnnualSalary[salaryGap$Currency == "USD"]  * chageUSDEUR

salaryGap$MenAverageAnnualSalaryEUR[salaryGap$Currency == "UKP"] <-
  salaryGap$MenAverageAnnualSalary[salaryGap$Currency == "UKP"] * 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

Contem el nombre de registres:

```
# Total
nrow(salaryGap)
```

```
## [1] 379
```

```
# Registres RU
nrow(salaryGapUK)
```

```
## [1] 237
```

```
# Registres EUA
nrow(salaryGapUS)
```

```
## [1] 142
```

Mostrem els resums de les variables numèriques en EUR

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2692   21810   30075   31841   39654   90003
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3766   26679   36152   39465   48205  106281
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -3921    3966    6559    7624   10066   38692
```

```
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)
```

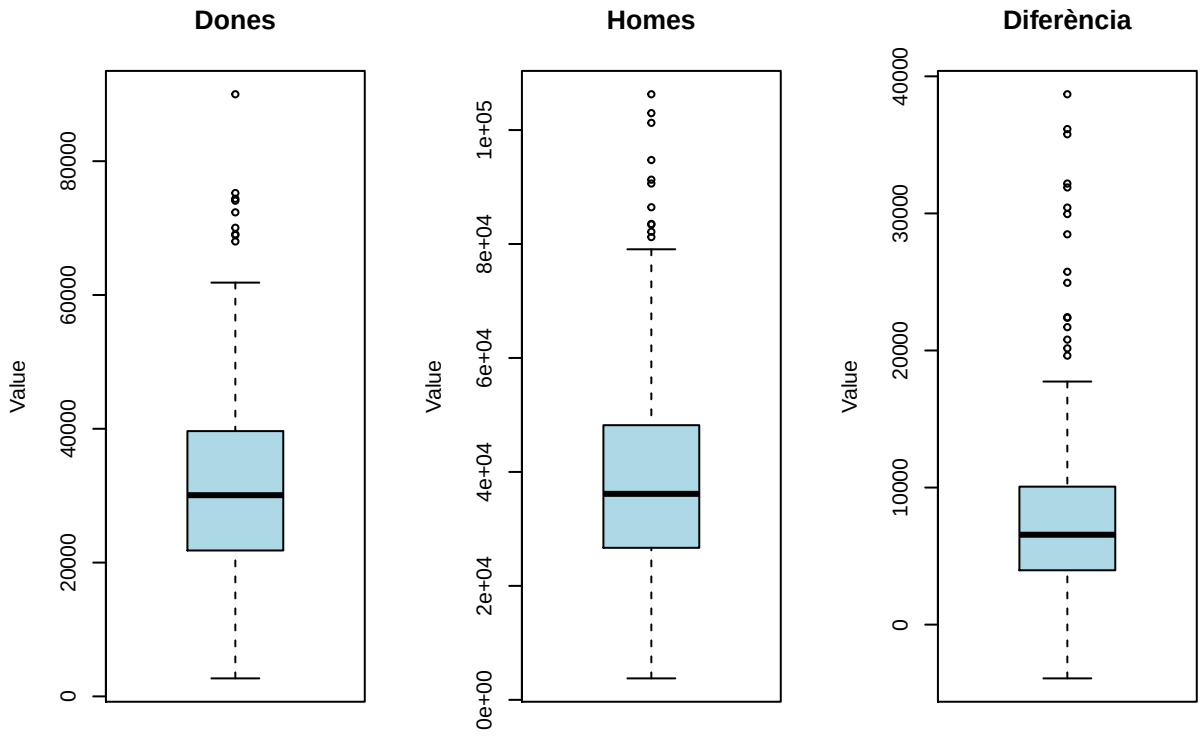
```
##               Admin        Construction Protective services
##                   3                   3                   3
##         Accountants        Arts & media              Bakers
##                   2                   2                   2
##    Care & education            Cashiers     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] 75238.80 90002.64 68945.24 72352.28 74387.04 70049.07 74093.76 68006.25
## [9] 69086.16
```

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

```
##  [1]  86453.64  94734.64  83519.80  90617.80 102968.32  82147.52  83425.16
##  [8] 106280.72  91289.25  81217.89 101264.67
```

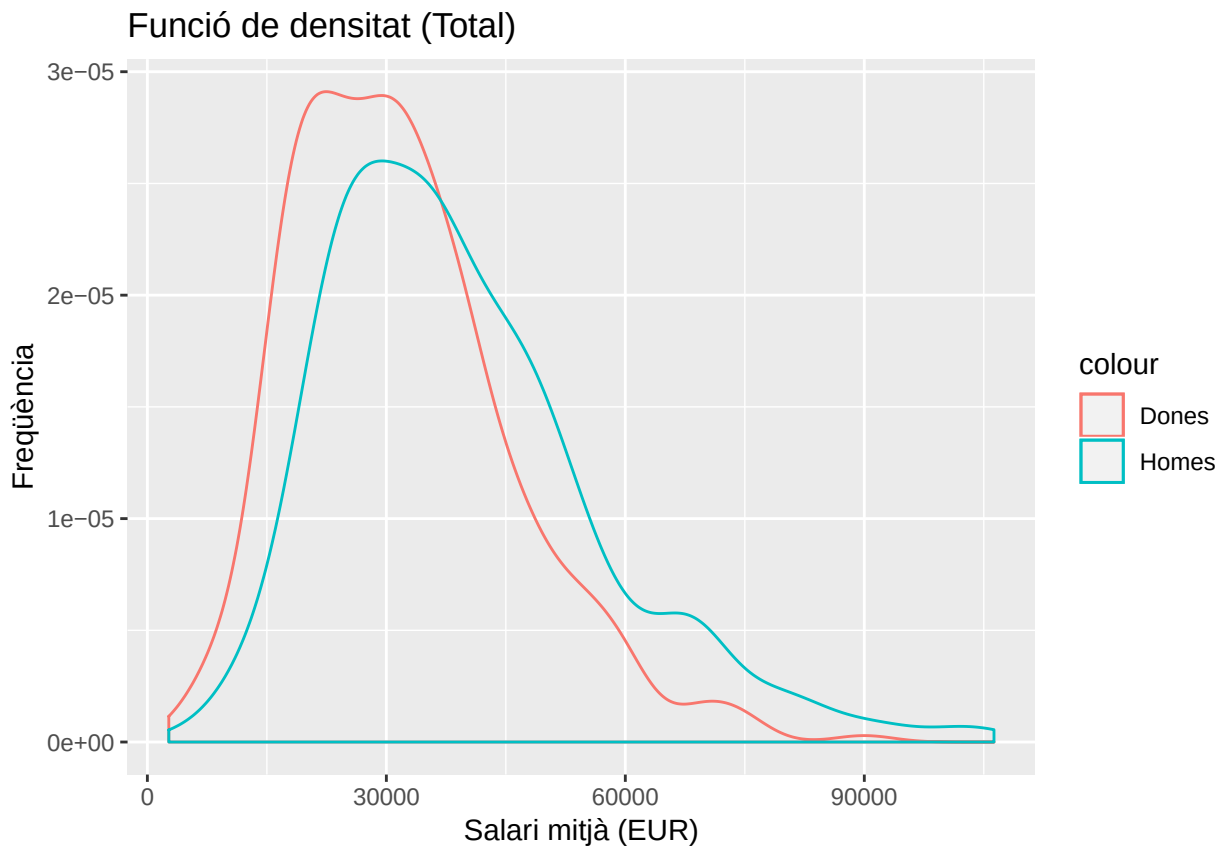
```
boxplot.stats(salaryGap$salaryGapEUR)$out
```

```
##  [1] 24937.64 35773.92 36152.48 22429.68 19623.24 29953.56 22382.36
##  [8] 25742.08 31893.68 20158.32 38691.90 21702.33 20793.24 30424.68
## [15] 28475.46 32178.51
```

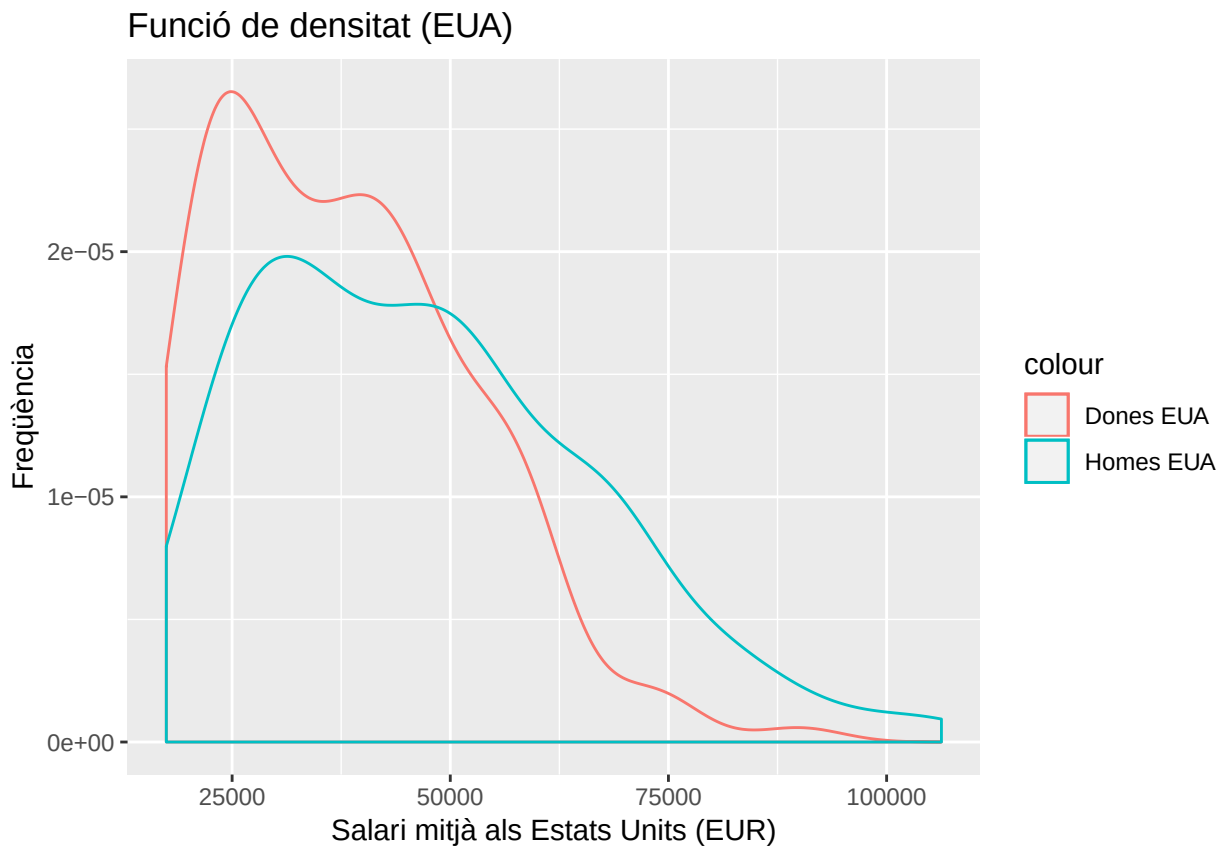
A la gràfica de densitat podem comprovar com les dades dels homes estan desplaçades a la dreta; això vol dir que trobem més valors masculins en el rang de salaris alts. També es pot veure com el pic de valors femenins és més alt i més a l'esquerra; això vol dir que hi ha moltes dones que guanyen poc.

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

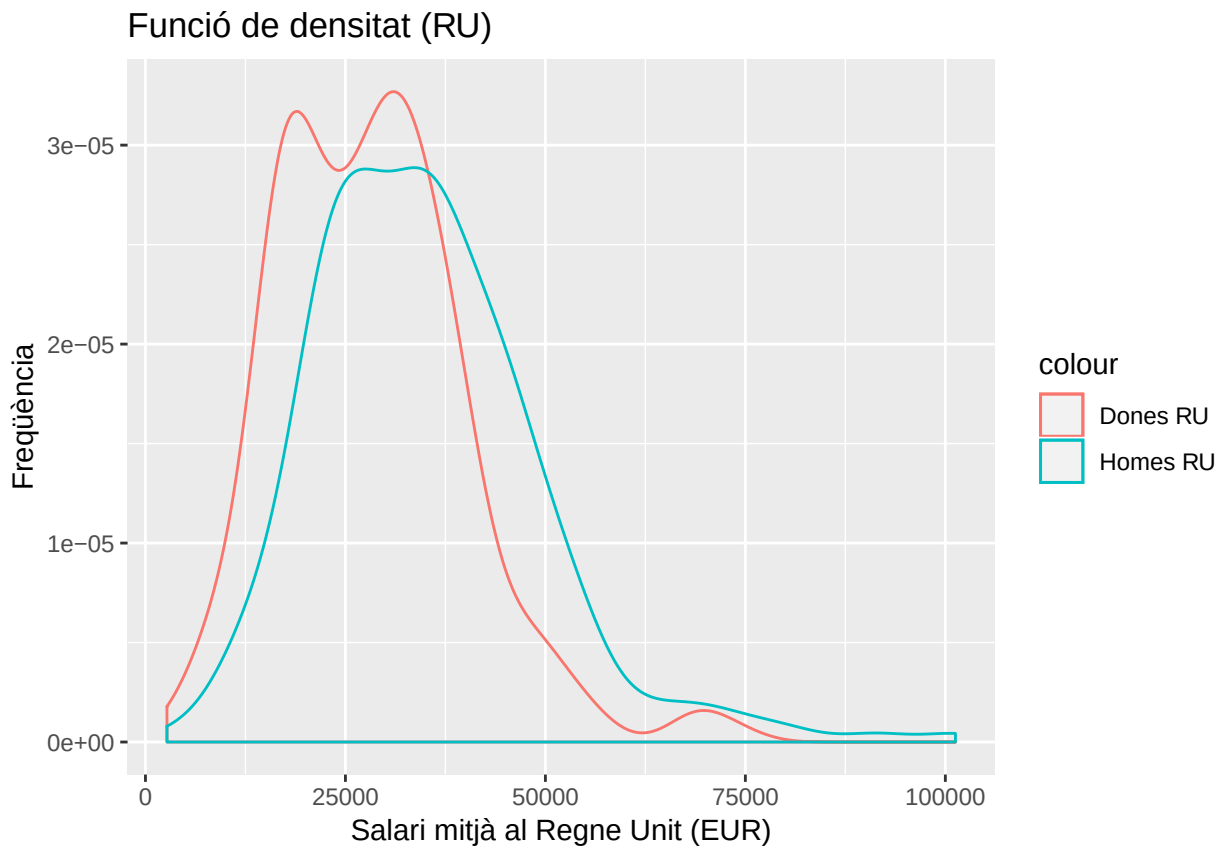
```
ggplot(salaryGap) +
  xlab("Salari mitjà (EUR)") + ylab("Freqüència") + ggtitle("Funció de densitat (Total)") +
  geom_density(aes(x = WomenAverageAnnualSalaryEUR, color = "Dones")) +
  geom_density(aes(x = MenAverageAnnualSalaryEUR, color = "Homes"))
```



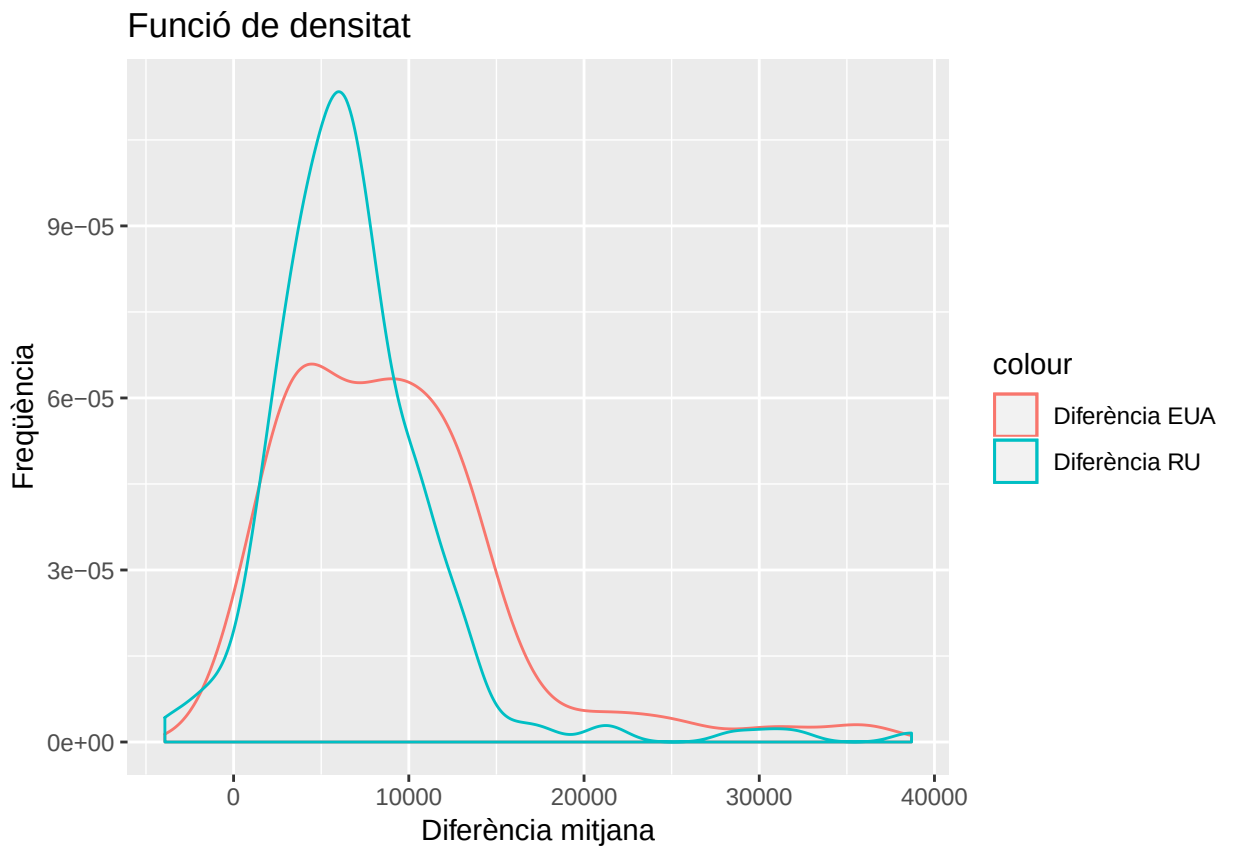

```
ggplot(salaryGapUS) +
  xlab("Salari mitjà als Estats Units (EUR)") + ylab("Freqüència") + ggtitle("Funció de densitat (EUA)")
  geom_density(aes(x = WomenAverageAnnualSalaryEUR, color = "Dones EUA")) +
  geom_density(aes(x = MenAverageAnnualSalaryEUR, color = "Homes EUA"))
```




```
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 10 casos:

```
salaryGapWM <- filter(salaryGap, WomenAverageAnnualSalary >= MenAverageAnnualSalary)
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')
      , row.names = TRUE)
```

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

	Ocupació	Categoria	País	Moneda	Salari dones	Salari homes
1	Stock clerks	sales & serving others	US	USD	24322.48	23849.28
2	Health technicians	science, tech & engineering	US	USD	29243.76	29243.76
3	Senior police	law & justice	UK	UKP	70049.07	67510.17
4	Traffic wardens	law & justice	UK	UKP	23426.91	21797.10
5	Valeters	sales & serving others	UK	UKP	17764.11	17119.44
6	Drivers	sales & serving others	UK	UKP	48897.81	47382.66
7	Train & tram drivers	sales & serving others	UK	UKP	56653.74	56093.31
8	Welfare officers	sales & serving others	UK	UKP	34011.90	30091.23
9	Town planners	science, tech & engineering	UK	UKP	34144.11	32438.25
10	IT directors	senior managers & execs	UK	UKP	74093.76	70451.55

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

El dashboard es pot trobar a:

https://my.visme.co/projects/vdjxvjdd-visualitzacio-de-dades

El resum de les operacions que s'han fet per poder mostrar les dades al dasboard són:

```
# Diferències totals
nrow(filter(salaryGap, salaryGapEUR <= 0))
```

```
## [1] 10
```

```
nrow(filter(salaryGap, salaryGapEUR > 0 & salaryGapEUR <= 5000))
```

```
## [1] 122
```

```
nrow(filter(salaryGap, salaryGapEUR > 5000 & salaryGapEUR <= 10000))
```

```
## [1] 151
```

```
nrow(filter(salaryGap, salaryGapEUR > 10000 & salaryGapEUR <= 15000))
```

```
## [1] 71
```

```
nrow(filter(salaryGap, salaryGapEUR > 15000))
```

```
## [1] 25
```

```
# Diferències als EUA
nrow(filter(salaryGapUS, salaryGapEUR <= 0))
```

```
## [1] 2
```

```
nrow(filter(salaryGapUS, salaryGapEUR > 0 & salaryGapEUR <= 5000))
```

```
## [1] 44
```

```
nrow(filter(salaryGapUS, salaryGapEUR > 5000 & salaryGapEUR <= 10000))
```

```
## [1] 44
```

```
nrow(filter(salaryGapUS, salaryGapEUR > 10000 & salaryGapEUR <= 15000))
```

```
## [1] 36
```

```
nrow(filter(salaryGapUS, salaryGapEUR > 15000))
```

```
## [1] 16
```

```
# Diferències as RU
nrow(filter(salaryGapUK, salaryGapEUR <= 0))
```

```
## [1] 8
```

```
nrow(filter(salaryGapUK, salaryGapEUR > 0 & salaryGapEUR <= 5000))
```

```
## [1] 78
```

```
nrow(filter(salaryGapUK, salaryGapEUR > 5000 & salaryGapEUR <= 10000))
```

```
## [1] 107
```

```r
nrow(filter(salaryGapUK, salaryGapEUR > 10000 & salaryGapEUR <= 15000))
```

```
## [1] 35
```

```r
nrow(filter(salaryGapUK, salaryGapEUR > 15000))
```

```
## [1] 9
```

```r
# Dades de les categories
salaryGapCategory <-
filter(salaryGap, tolower(as.character(Occupation)) == tolower(as.character(Category)))
salaryGapCategory <- select (salaryGapCategory,
                        Category, Country,   Currency,
                        WomenAverageAnnualSalaryEUR, MenAverageAnnualSalaryEUR
                        , salaryGapEUR, PayGapAsAPercentage)
salaryGapCategory <- arrange (salaryGapCategory, Country, Category)
kable(salaryGapCategory, caption = "Categories"
      ,col.names = c("Categoria","País", "Moneda", "Salari dones","Salari homes"
                     , "Diferència", "Diferència %")
      ,align = c('r', 'r', 'r', 'r', 'r', 'r')
      , row.names = TRUE)
```

Table 2: Categories

	Categoria	País	Moneda	Salari dones	Salari homes	Diferència	Diferència %
1	admin & organisation	UK	UKP	23844.60	29613.87	5769.27	0.2369
2	care & education	UK	UKP	28478.97	34602.75	6123.78	0.2110
3	creative & media	UK	UKP	30859.92	36948.60	6088.68	0.2119
4	law & justice	UK	UKP	33783.75	37200.15	3416.40	0.1193
5	manual work	UK	UKP	19320.21	26358.93	7038.72	0.3809
6	sales & serving others	UK	UKP	21608.73	26635.05	5026.32	0.2812
7	science, tech & engineering	UK	UKP	32556.42	39996.45	7440.03	0.2537
8	senior managers & execs	UK	UKP	37640.07	48002.76	10362.69	0.2823
9	admin & organisation	US	USD	37983.40	46708.48	8725.08	0.2102
10	care & education	US	USD	39982.67	48421.10	8438.43	0.2017
11	creative & media	US	USD	42114.80	51878.19	9763.39	0.2311
12	law & justice	US	USD	28756.91	36804.04	8047.13	0.2622
13	manual work	US	USD	22415.12	27720.42	5305.30	0.2382
14	sales & serving others	US	USD	31869.11	38856.09	6986.98	0.2155
15	science, tech & engineering	US	USD	46429.11	55561.87	9132.76	0.2046
16	senior managers & execs	US	USD	51763.53	71386.77	19623.24	0.3779

```r
# Top diferència
salaryGapTop <-
  filter(salaryGap, tolower(as.character(Occupation)) != tolower(as.character(Category)))
salaryGapTop <- select (salaryGapTop,
                        Occupation, Category, Country,
                        WomenAverageAnnualSalaryEUR,
                        MenAverageAnnualSalaryEUR, salaryGapEUR, PayGapAsAPercentage)
salaryGapTop <- arrange (salaryGapTop, desc(salaryGapEUR))
salaryGapTop <- head(salaryGapTop)
kable(salaryGapTop, caption = "Top diferències"
      ,col.names = c("Ocupació", "Categoria","País", "Salari dones","Salari homes",
                     "Diferència", "Diferència %")
```

```
      ,align = c('r', 'r', 'r', 'r', 'r', 'r')
      , row.names = TRUE)
```

Table 3: Top diferències

	Ocupació	Categoria	País	Salari dones	Salari homes	Diferència	Diferència %
1	Medical practitioners	care & education	UK	52597.35	91289.25	38691.90	0.7356
2	Legal occupations	law & justice	US	47367.32	83519.80	36152.48	0.7632
3	Doctors & surgeons	care & education	US	58960.72	94734.64	35773.92	0.6067
4	CEOs	senior managers & execs	UK	69086.16	101264.67	32178.51	0.4658
5	Chief executives	senior managers & execs	US	74387.04	106280.72	31893.68	0.4288
6	Bank managers	senior managers & execs	UK	43440.93	73865.61	30424.68	0.7004

```
salaryGapTop <-
filter(salaryGap, tolower(as.character(Occupation)) != tolower(as.character(Category)))
salaryGapTop <- select (salaryGapTop,
                        Occupation, Category, Country,
                        WomenAverageAnnualSalaryEUR, MenAverageAnnualSalaryEUR,
                        salaryGapEUR, PayGapAsAPercentage)
salaryGapTop <- arrange (salaryGapTop, desc(PayGapAsAPercentage))
salaryGapTop <- head(salaryGapTop)
kable(salaryGapTop, caption = "Top diferències en percentatge"
      ,col.names = c("Ocupació", "Categoria","País", "Salari dones","Salari homes",
                     "Diferència", "Diferència %")
      ,align = c('r', 'r', 'r', 'r', 'r', 'r')
      , row.names = TRUE)
```

Table 4: Top diferències en percentatge

	Ocupació	Categoria	País	Salari dones	Salari homes	Diferència	Diferència %
1	Cleaning occupations	sales & serving others	UK	7305.48	13672.62	6367.14	0.8716
2	Machine operatives	science, tech & engineering	UK	16162.38	29577.60	13415.22	0.8300
3	Legal occupations	law & justice	US	47367.32	83519.80	36152.48	0.7632
4	Medical practitioners	care & education	UK	52597.35	91289.25	38691.90	0.7356
5	Bank managers	senior managers & execs	UK	43440.93	73865.61	30424.68	0.7004
6	Metalwork	manual work	UK	17636.58	29886.48	12249.90	0.6946

```
salaryGapTop <-
filter(salaryGap, tolower(as.character(Occupation)) != tolower(as.character(Category)))

salaryGapTop <-
  filter(salaryGapTop, salaryGapEUR < 0)

salaryGapTop <- select (salaryGapTop,
                        Occupation, Country,
                        WomenAverageAnnualSalaryEUR, MenAverageAnnualSalaryEUR,
                        salaryGapEUR, PayGapAsAPercentage)

salaryGapTop <- arrange (salaryGapTop, salaryGapEUR)
kable(salaryGapTop, caption = "Diferències en favor de les dones"
      ,col.names = c("Ocupació", "País", "Salari dones","Salari homes",
```

```
                    "Diferència", "Diferència %")
      ,align = c('r', 'r', 'r', 'r', 'r', 'r')
      , row.names = TRUE)
```

Table 5: Diferències en favor de les dones

	Ocupació	País	Salari dones	Salari homes	Diferència	Diferència %
1	Welfare officers	UK	34011.90	30091.23	-3920.67	-0.1153
2	IT directors	UK	74093.76	70451.55	-3642.21	-0.0492
3	Senior police	UK	70049.07	67510.17	-2538.90	-0.0362
4	Town planners	UK	34144.11	32438.25	-1705.86	-0.0500
5	Traffic wardens	UK	23426.91	21797.10	-1629.81	-0.0696
6	Drivers	UK	48897.81	47382.66	-1515.15	-0.0310
7	Valeters	UK	17764.11	17119.44	-644.67	-0.0363
8	Train & tram drivers	UK	56653.74	56093.31	-560.43	-0.0099
9	Stock clerks	US	24322.48	23849.28	-473.20	-0.0195

Finalment, desem el dataset:

```
write.csv(salaryGap, file = "salaryGapFinal.csv")
```