PRAC2 Tipologia i cicle de vida de les dades

Autor: Edgar Pardo - Pau Campaña

May 2020

Contents

```
library(corrplot)
library(RColorBrewer)
library(ggplot2)
library(reshape)
library(caret)
library(devtools)
library(ggbiplot)
library(dplyr)
library(ISLR)
library(nortest)
```

Descripció

El conjunt de dades que s'analitza en aquesta pràctica tracta sobre el salari i diferents característiques de les persones. Està disponible al següent enllaç de Kaggle https://www.kaggle.com/pcampana/wagepeople

Objectiu de l'anàlisi

A partir d'aquest conjunt de dades es planteja la problemàtica de determinar quines variables influeixen més sobre el salari d'una persona. Per fer-ho, es farà ús de diferents proves estadístiques. Es farà un anàlisi de correlació on es mirarà quines variables influeixen més en el salari d'una persona. Calcularem l'interval de confiança del salari per diferents grups de persones, i veurem quins grups guanyen més depenent de les seves característiques. A partir d'un contrast d'hipòtesi, veurem si les persones amb assegurança privada cobren més que les que no en tenen.

Amb aquestes proves pretenem estudiar a fons el conjunt de dades, i poder acabar concloent què fa que una persona cobri més.

Descripció del dataset

Llegim el conjunt de dades que es troba en el fitxer dataset_original.

```
dataset <- read.csv("../data/dataset_original.csv")
```

Si inspeccionem les dades podrem veure per una banda el tamany del dataset, i per altra banda el tipus de cadascuna de les variables,

```
files = dim(dataset)[1] #files del dataframe
columnes = dim(dataset)[2] #columnes del dataframe
cat('Tamany del dataset:\n
    - Files: ', files ,'\n
    - Columnes: ', columnes ,'\n'
    )
```

```
## Tamany del dataset:
##
##     - Files:  3000
##
##     - Columnes:  12
```

```
sapply(dataset, function(x) class(x))
```

```
##          X       year        age     maritl       race  education
##  "integer"  "integer"  "integer"   "factor"   "factor"   "factor"
##     region   jobclass     health health_ins    logwage       wage
##   "factor"   "factor"   "factor"   "factor"  "numeric"  "numeric"
```

Veiem que el tipus de dades assignat per part del propi R a cadascun dels atributs, es correspon amb la pròpia naturalesa i domini d'aquestes. Tenim un dataframe de 3000 treballadors amb 12 atributs per cadascun d'ells. Els atributs són de tres tipus:

- int
- Factor
- num

Les variables numèriques són:

- X: identificador de la persona
- year: fa referència a l'any en el qual es va recollir la informació del salari corresponent a aquella fila.
- age: edat del treballador en el moment en què es va recollir la informació.
- logwage: registre del salari del treballador
- wage: salari brut del treballador per 1000 $

Les variables de tipus factor:

- maritl: fa referència a l'estat civil del treballador. Pot prendre diversos valors: 1. Never Married 2. Married 3. Widowed 4. Divorced and 5. Separated.
- race: fa referència a la raça del treballador. Pot prendre els següents valors: 1. White 2. Black 3. Asian and 4. Other.
- education: fa referència al nivell d'estudis del treballador. Pot prendre els següents valors: 1. < HS Grad 2. HS Grad 3. Some College 4. College Grad and 5.Advanced Degree
- region: fa referència a la regió on viu el treballador. Només pren un valor: mid-atlantic
- jobclass: fa referència al tipus de treball. Pot prendre els següents valors: 1. Industrial and 2. Information.
- health: fa referència a l'estat de salut. Pot prendre els següents valors: 1. <=Good and 2. >=Very Good
- health_ins: fa referència a si el treballador té sanitat privada o no. Pot prendre els següents valors: 1. Yes and 2. No.

Selecció de les dades d'interès

Els atributs que trobem en el dataset fan referència a característiques de les persones, que seran d'utilitat per la realització de l'anàlisi. Tot i això, trobem un camp en el dataset que no aporta cap informació útil. És l'atribut 'X' que és l'identificador de la persona. Com que no és una dada d'interès, l'eliminarem.

```
dataset <- dataset[, -(1:1)]
```

Neteja de dades

Valors nuls

Un cop tenim el conjunt de dades insertat, inspeccionem si existeixen valors nuls. Per fer-ho, inspeccionarem cada columna del dataset buscant valors buits.

```
colSums(is.na(dataset))
```

```
##       year        age     maritl       race  education     region
##          0          0          0          0          0          0
##   jobclass     health health_ins    logwage       wage
##          0          0          0          0          0
```

Podem observar que no hi ha valors buits, pel que no és necessari fer cap procés de neteja al conjunt de dades referent a valors nuls.

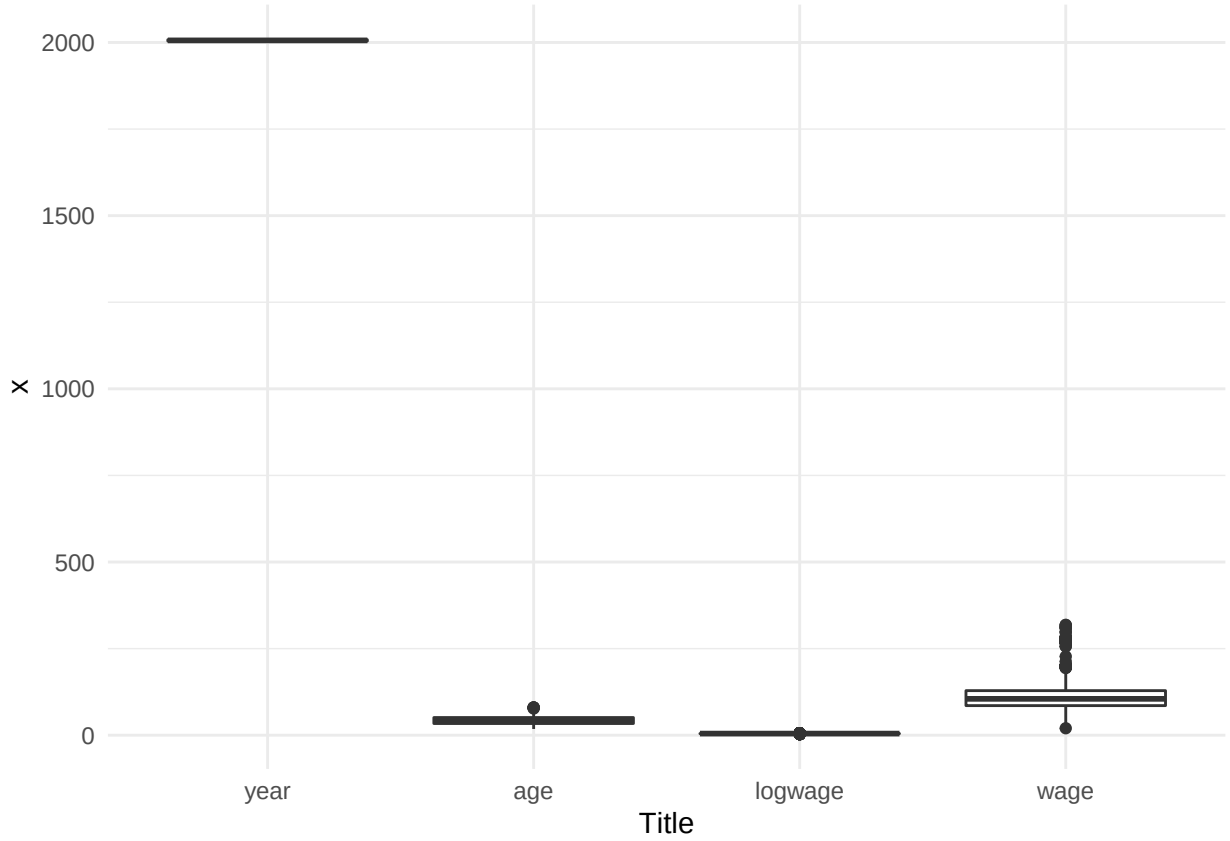
Valors extrems

Els valors extrems són aquells que són molt distants a la resta. Per estudiar els valors extrems, ens centrarem en les variables numèriques del dataset. Visualitzem els valors extrems amb gràfiques boxplot per tal de veure-ho representat de manera gràfica.

```
x <- melt(dataset)
```

```
## Using maritl, race, education, region, jobclass, health, health_ins as id variables
```

```
plt <- ggplot(data = x, aes(x = variable, y = value))
plt + geom_boxplot() + theme_minimal() + labs(x = "Title", y = "x")
```



S'observa que hi ha diferents valors extrems. Observarem quins valors prenen els outliers per tal de poder decidir que fer amb ells.

```
cat('Outliers year:\n')
```

```
## Outliers year:
```

```
boxplot.stats(dataset$year)$out
```

```
## integer(0)
```

```
cat('Outliers age:\n')
```

```
## Outliers age:
```

```
boxplot.stats(dataset$age)$out
```

```
## [1] 80 80 80 80
```

```
cat('Outliers logwage:\n')
```

```
## Outliers logwage:
```

```
boxplot.stats(dataset$logwage)$out
```

```
##   [1] 5.626186 3.544068 3.812913 5.590618 5.606885 3.556303 5.591980
##   [8] 5.623217 3.544068 5.763128 5.546741 3.778151 3.477121 3.778151
##  [15] 3.716003 5.641006 3.653405 5.606885 5.606885 3.041393 5.606885
##  [22] 5.590618 5.650820 3.133858 5.750441 3.730621 5.626900 5.626900
##  [29] 3.477121 5.607283 3.301030 5.606885 3.778151 5.641006 3.301030
##  [36] 5.626900 5.631919 5.590618 5.626900 3.602060 5.631084 5.641006
##  [43] 5.590618 5.590618 3.556303 5.606885 5.641006 5.626186 5.735190
##  [50] 5.641006 3.698970 5.763128 5.591980 5.606885 5.591980 5.742793
##  [57] 5.591980 5.606885 3.778151 5.606885 3.778151 3.477121 3.643453
##  [64] 3.230449 5.626186 5.606885 3.698970 5.623217 5.623217 5.590618
##  [71] 3.778151 5.623217 3.176091 5.623217 3.806180 5.626900 3.380211
##  [78] 5.623217 3.477121 5.633009 5.626186 3.544068 3.301030 3.431525
##  [85] 5.626900 5.546741 5.591980 3.000000 3.447158 3.602060 5.623217
##  [92] 5.626900 5.641006 3.698970 3.693727 5.590618 5.623217 5.591980
##  [99] 5.623217 5.626900 3.301030 3.544068 5.606885 5.591980 5.590618
## [106] 5.623217 5.626186 3.778151 3.698970 3.778151 3.602060 3.724276
## [113] 3.176091 3.147367 5.626900 5.626186 5.591980 5.626186 5.590618
## [120] 5.591980 5.735190 5.591980 5.590618 5.641006 5.701323 5.690330
```

```
cat('Outliers wage:\n')
```

```
## Outliers wage:
```

```
boxplot.stats(dataset$wage)$out
```

```
##   [1] 212.84235 200.54326 277.60142 195.67631 267.90109 200.54326 272.29478
##   [8] 200.54326 200.54326 268.26629 200.54326 276.77841 318.34243 256.40065
##  [15] 281.74597 272.29478 272.29478 272.29478 200.54326 267.90109 284.52474
##  [22] 314.32934 277.79948 277.79948 196.12527 200.54326 272.40323 200.54326
##  [29] 272.29478 281.74597 277.79948 200.54326 279.19750 193.86688 267.90109
##  [36] 198.35029 277.79948 278.96447 281.74597 196.12527 267.90109 267.90109
##  [43] 200.54326 272.29478 200.54326 281.74597 277.60142 309.57177 281.74597
##  [50] 200.54326 318.34243 198.35029 268.26629 272.29478 268.26629 311.93457
##  [57] 268.26629 272.29478 272.29478 277.60142 272.29478 200.54326 276.77841
##  [64] 276.77841 196.12527 267.90109 276.77841 276.77841 277.79948 276.77841
##  [71] 279.50178 277.60142 277.79948 256.40065 268.26629 200.54326 200.54326
##  [78]  20.08554 198.35029 227.45893 276.77841 277.79948 281.74597 267.90109
##  [85] 276.77841 268.26629 196.12527 276.77841 277.79948 272.29478 268.26629
##  [92] 267.90109 276.77841 277.60142 196.12527 277.79948 277.60142 268.26629
##  [99] 277.60142 267.90109 200.54326 268.26629 309.57177 200.54326 200.54326
## [106] 268.26629 267.90109 281.74597 299.26298 295.99125
```

Observem que els atributs age, logwage i wage tenen outliers. Mirant els diferents outliers, veiem que són valors possibles, i que no es tracta d'errors del dataset. Per exemple, pel cas dels anys. Veiem que 80 anys apareix com a valor extrem, tot i que per context es veu que és un valor possible dins del dataset. De la mateixa manera passa pels atributs logwage i wage. Per tant, al tractar-se de valors que es poden donar perfectament, s'ha decidit que el tractament dels valors extrems serà deixar-los tal i com estan.

Extracció de les dades

Un cop hem estudiat els valors extrems i nuls, i hem eliminat les columnes innecessàries per aquest estudi, podem desar en un nou fitxer denominat *dataset_clean* el conjunt de dades resultant.

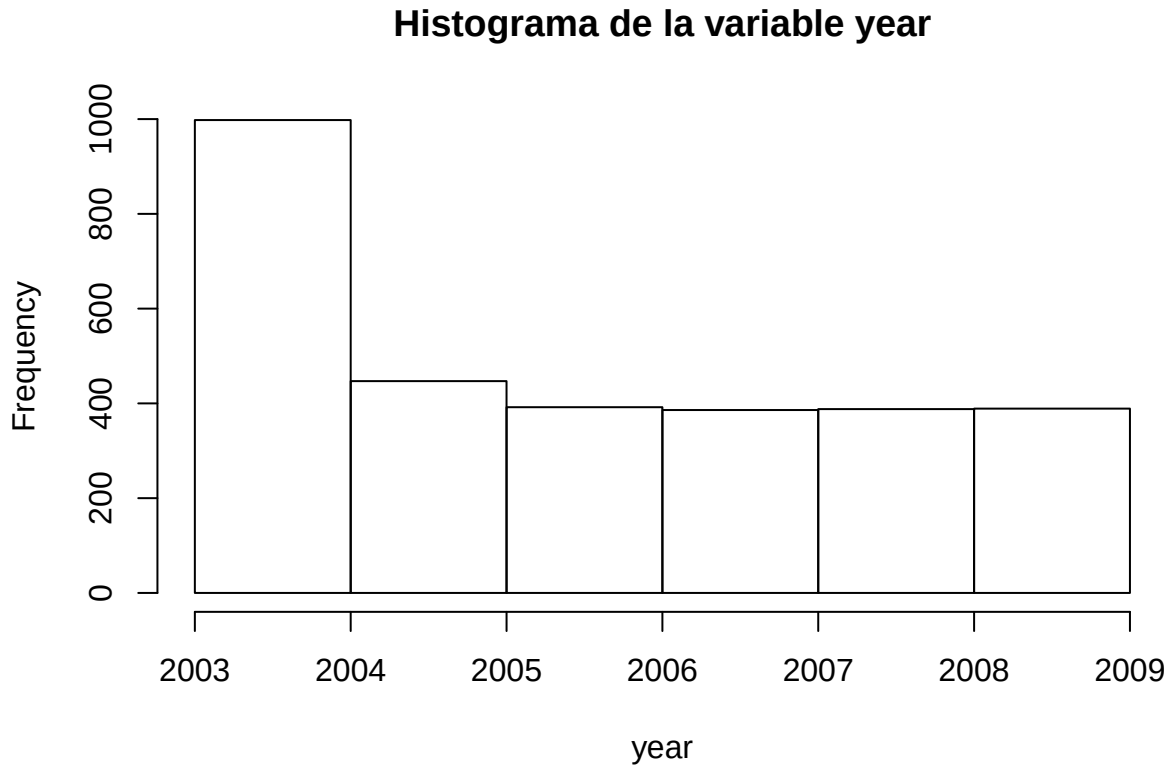
```
write.csv(dataset, "../data/dataset_clean.csv")
```

Anàlisi de les dades

Començarem inspeccionant els diferents atributs que tenim al dataset. Un cop carregades les dades, és moment de fer una descriptiva numèrica de les dades. Ho mirarem atribut a atribut:

year

```
hist(dataset$year, xlab="year", main="Histograma de la variable year", breaks=6)
```



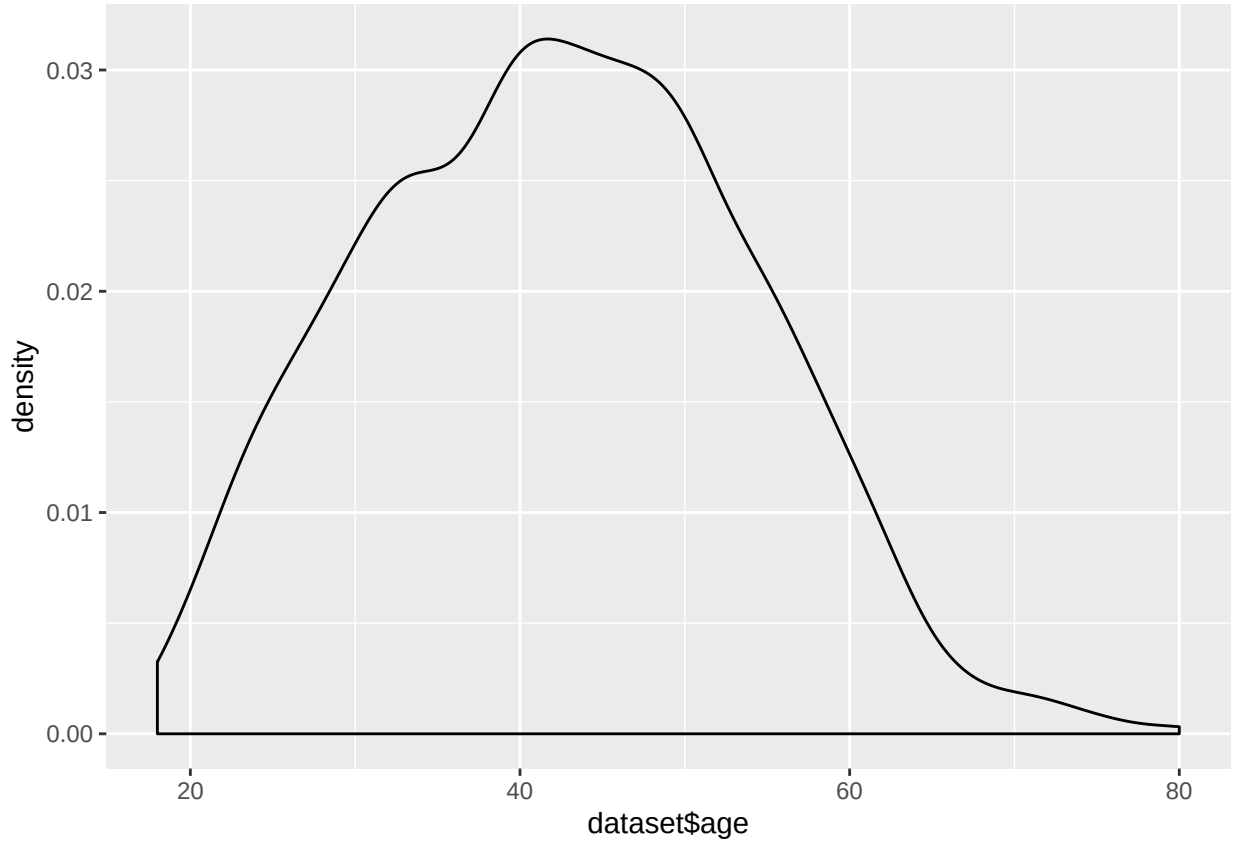

```
summary(dataset$year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2003    2004    2006    2006    2008    2009
```

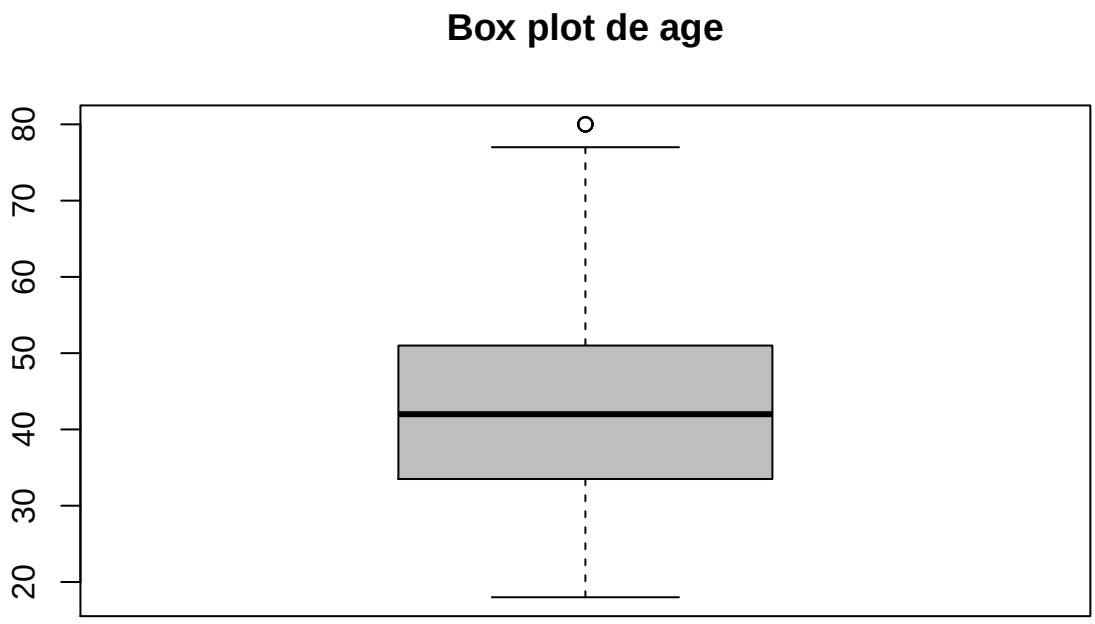
És una variable numèrica discreta on la majoria de tuples es concentren en l'any 2003 (l'any que es van recollir les dades). El valor màxim és 2009 i mínim 2003.

age

```
library(ggplot2)
ggplot(mapping= aes(x=dataset$age))+ geom_density()
```




```
boxplot(dataset$age,main="Box plot de age", col="gray")
```



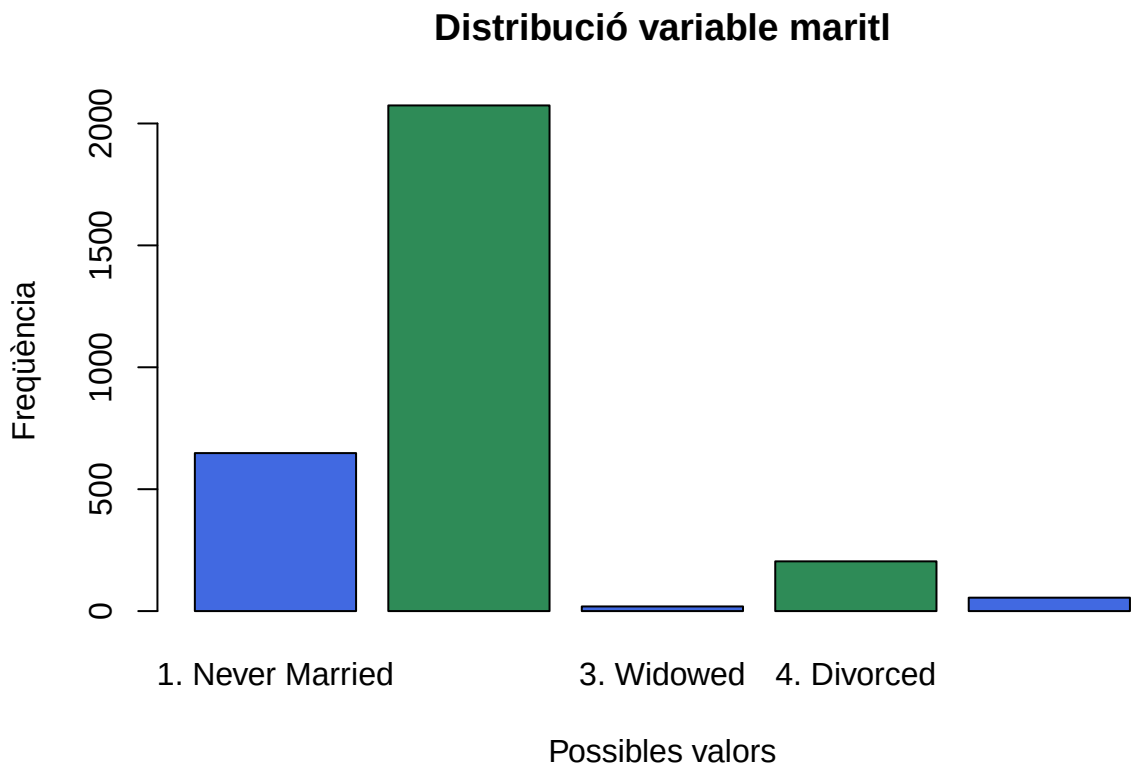

```
summary(dataset$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   33.75   42.00   42.41   51.00   80.00
```

És una variable numèrica podríem dir continua amb valor mínim 18 i valor màxim 80. Els valors es concentren al voltant de l'interval 40-50 anys (mitjana=42.41). Sembla seguir una distribució normal.

maritl:

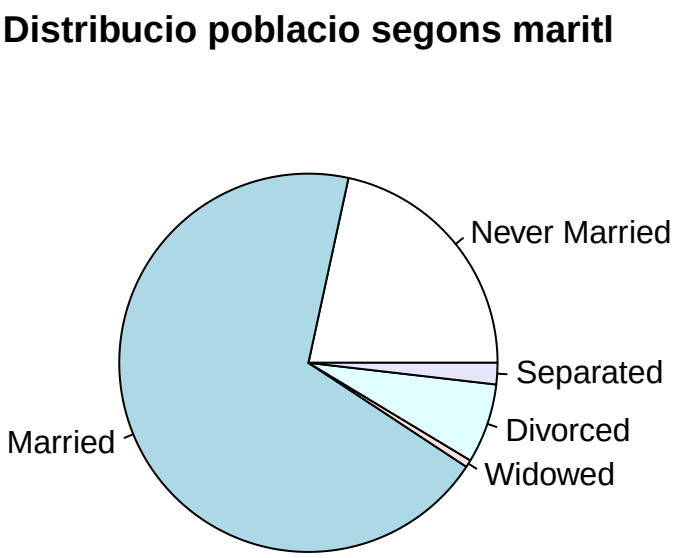
```
plot(x = dataset$maritl, main = "Distribució variable maritl",xlab = "Possibles valors", ylab = "Freqüèn
```




```
table(dataset$maritl)
```

```
##
## 1. Never Married       2. Married       3. Widowed      4. Divorced
##              648             2074               19              204
##     5. Separated
##               55
```

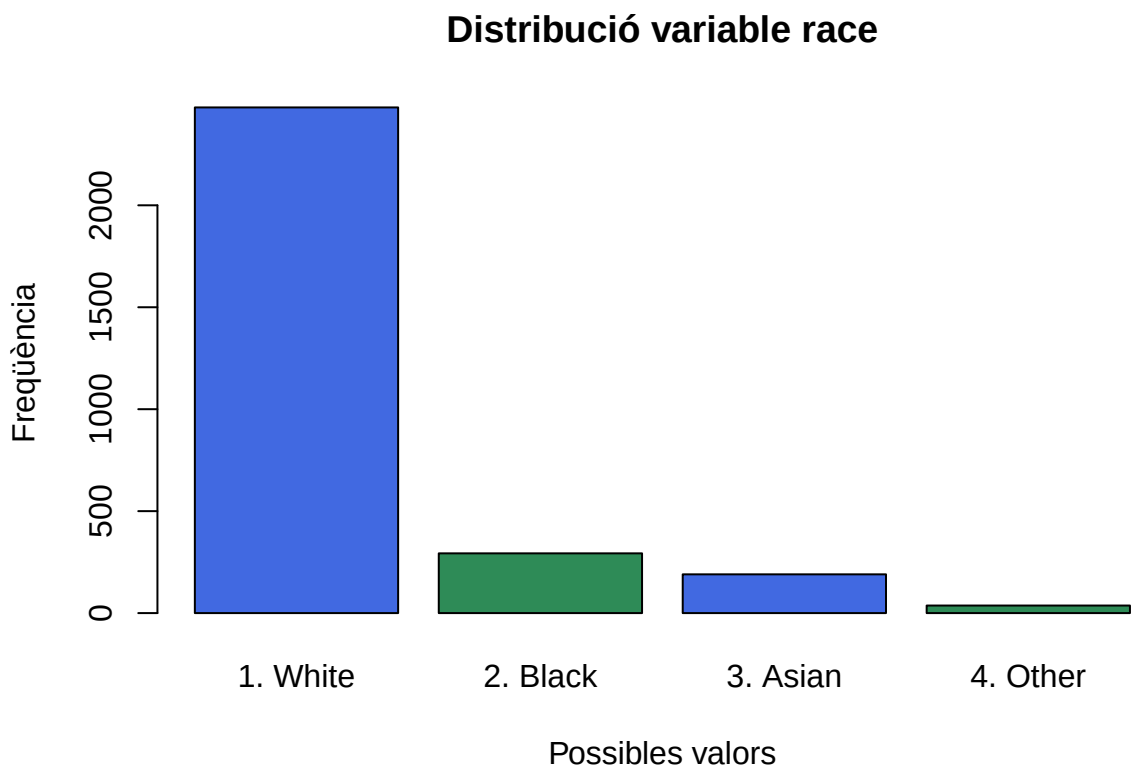
```
slices <- c(table(dataset$maritl))
lbls <- c("Never Married", "Married", "Widowed", "Divorced","Separated")
pie(slices, labels = lbls, main="Distribucio poblacio segons maritl")
```



Variable de tipus factor que pot prendre els cinc valors que es veuen al gràfic circular i explicats en el primer exercici. Observem com la majoria dels treballadors presenten l'estat civil casat.

race

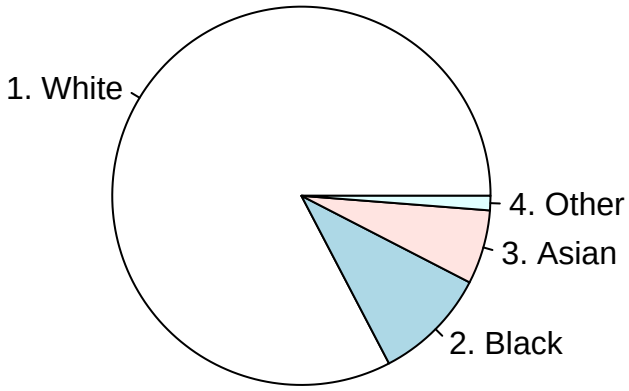
```
plot(x = dataset$race, main = "Distribució variable race",xlab = "Possibles valors", ylab = "Freqüència
```




```
table(dataset$race)
```

```
##
## 1. White 2. Black 3. Asian 4. Other
##     2480      293      190       37
```

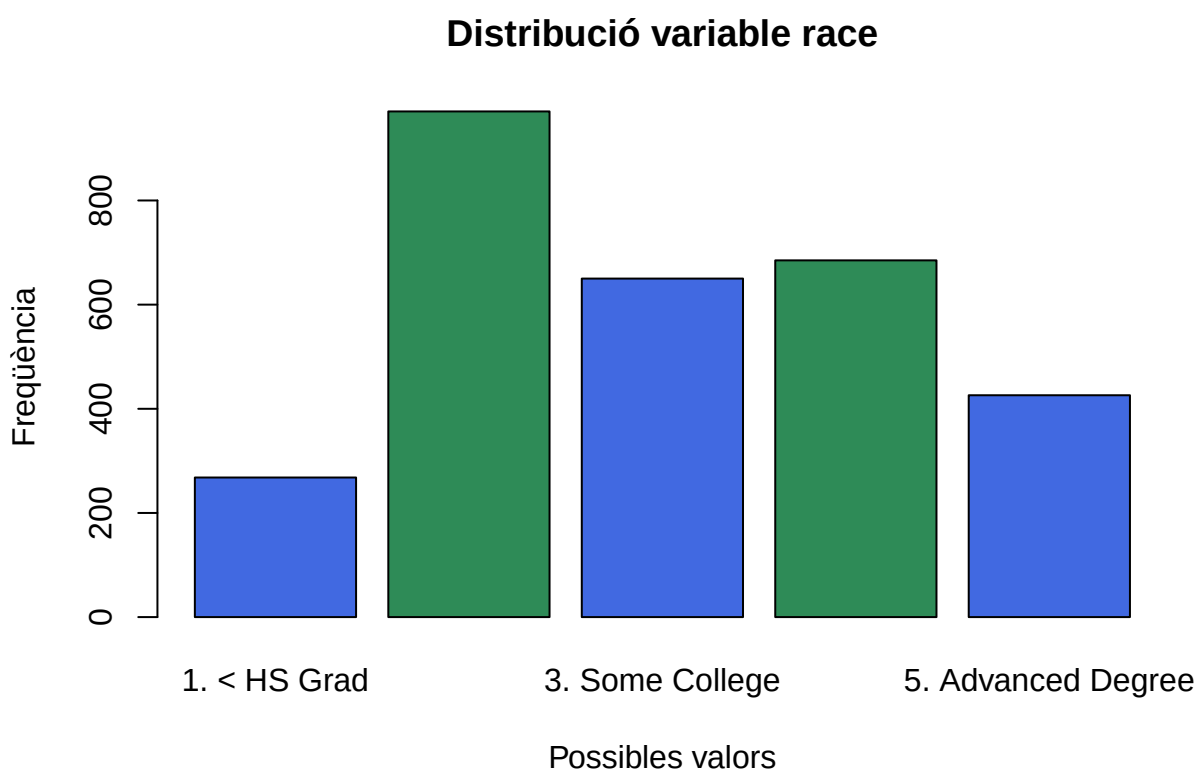
```
slices <- c(table(dataset$race))
lbls <- c("1. White", "2. Black", "3. Asian", "4. Other")
pie(slices, labels = lbls, main="Distribucio poblacio segons race")
```



Variable de tipus factor que pot prendre quatre valors i on la majoria dels treballadors són d'ètnia blanca (2480 persones).

education

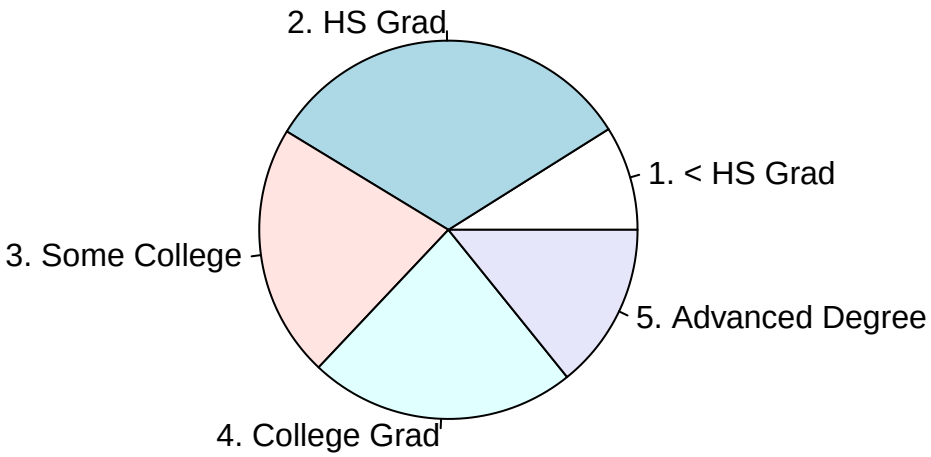
```
plot(x = dataset$education, main = "Distribució variable race",xlab = "Possibles valors", ylab = "Freqü
```




```
table(dataset$education)
```

```
##
##       1. < HS Grad          2. HS Grad    3. Some College
##                268                 971                 650
##    4. College Grad 5. Advanced Degree
##                685                 426
```

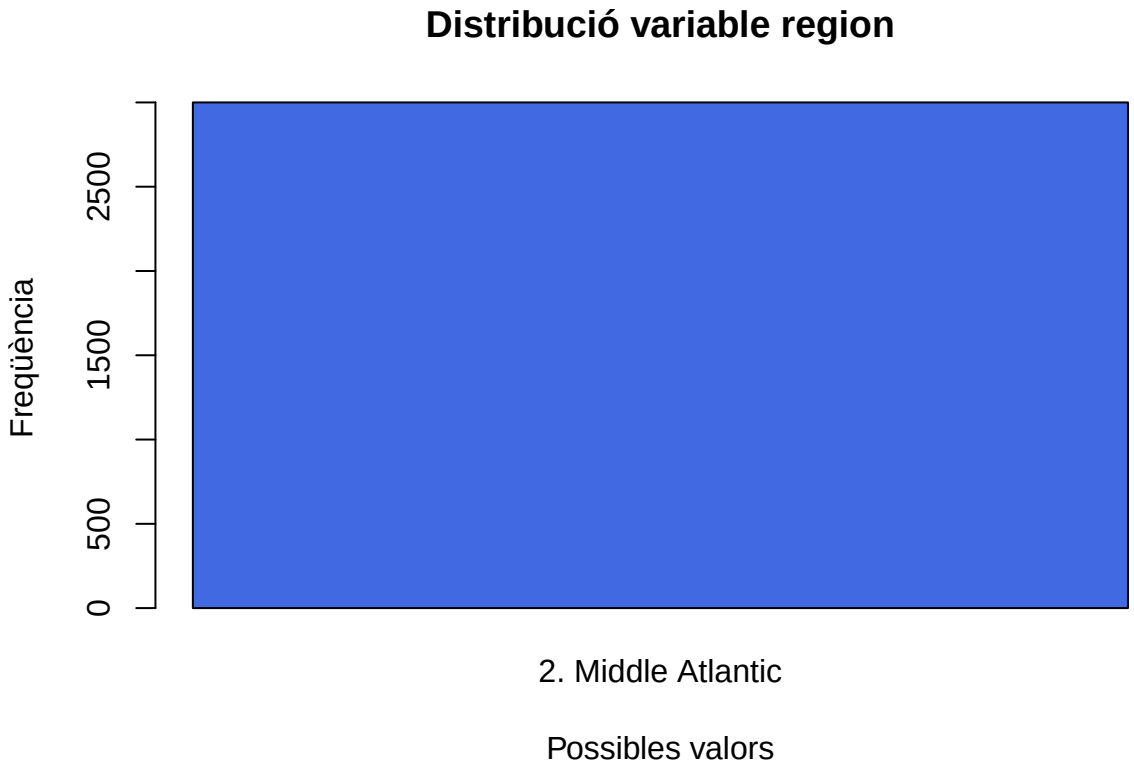
```
slices <- c(table(dataset$education))
lbls <- c("1. < HS Grad", "2. HS Grad", "3. Some College", "4. College Grad", "5. Advanced Degree")
pie(slices, labels = lbls, main="Distribucio poblacio segons education")
```



Variable de tipus factor que pot prendre cinc valors. En aquest cas, veiem que la mostra està més distribuïda entre les possibles opcions. El grup que més representat és el de les persones amb *HS Grad*, és a dir amb educació secundària.

region

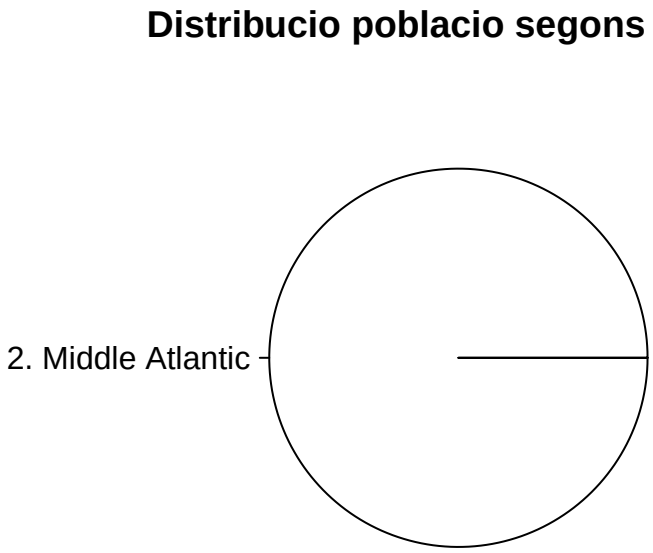
```
plot(x = dataset$region, main = "Distribució variable region",xlab = "Possibles valors", ylab = "Freqüèn
```




```
table(dataset$region)
```

```
##
## 2. Middle Atlantic
##               3000
```

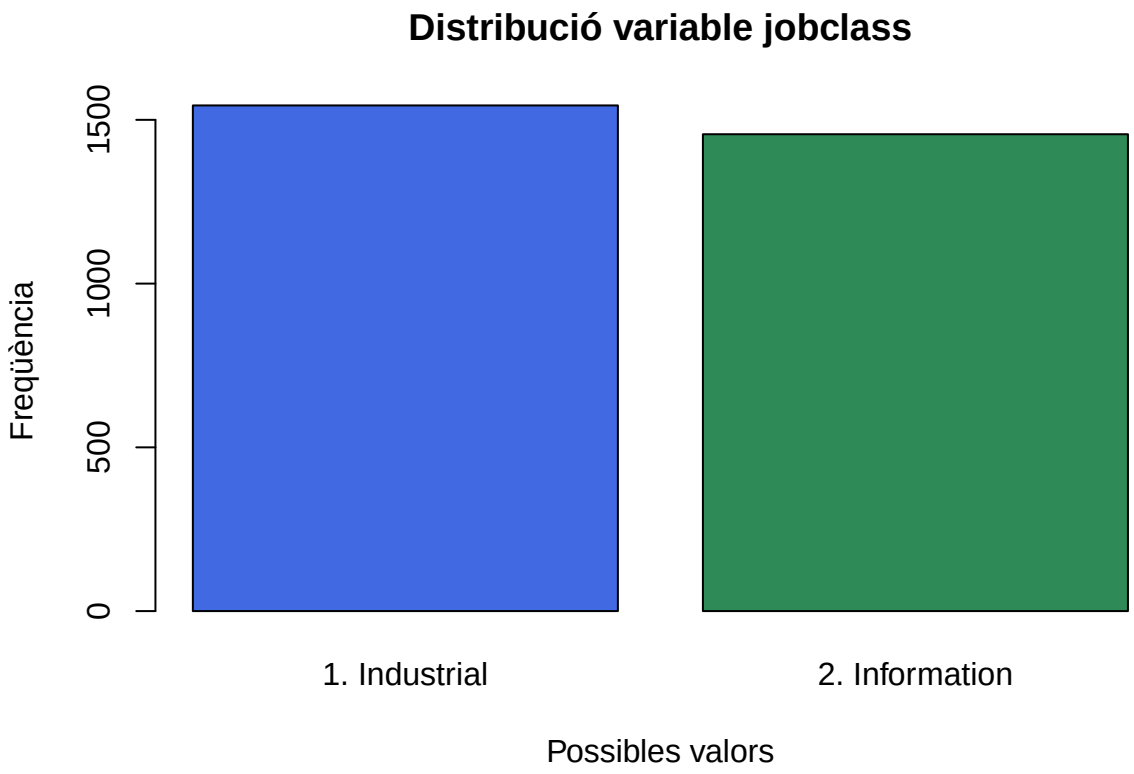
```
slices <- c(table(dataset$region))
lbls <- c("2. Middle Atlantic")
pie(slices, labels = lbls, main="Distribucio poblacio segons region")
```



Variable de tipus factor on tota la població pertany a una única regió: *Middle Atlantic*.

jobclass

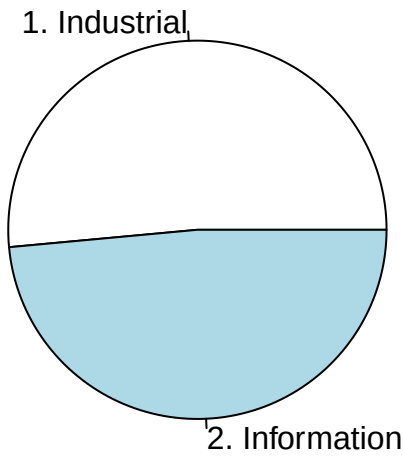
```
plot(x = dataset$jobclass, main = "Distribució variable jobclass",xlab = "Possibles valors", ylab = "Fr
```




```
table(dataset$jobclass)
```

```
##
##  1. Industrial 2. Information
##           1544           1456
```

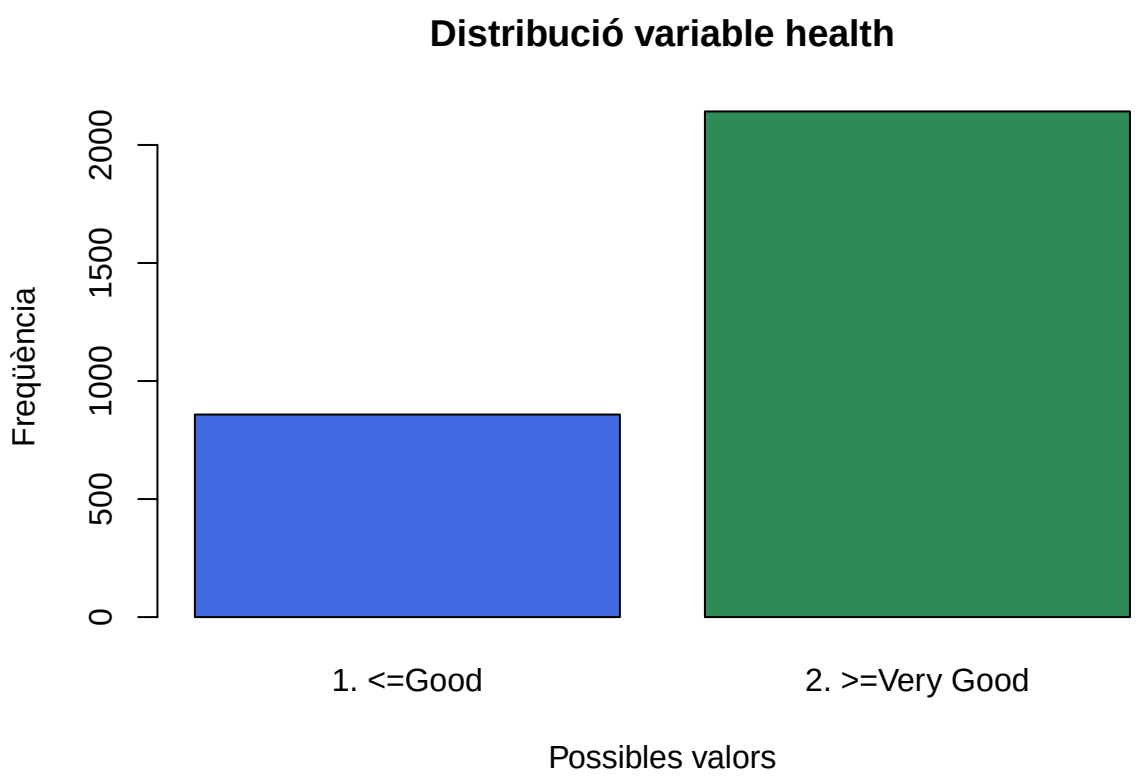
```
slices <- c(table(dataset$jobclass))
lbls <- c("1. Industrial","2. Information")
pie(slices, labels = lbls, main="Distribucio poblacio segons jobclass")
```



Variable de tipus factor que pot prendre dos valors: *industrial* i *information*. Els dos valors prenen quasi el mateix nom de tuples, però hi ha una lleugera diferència (88) en favor de *industrial*.

health

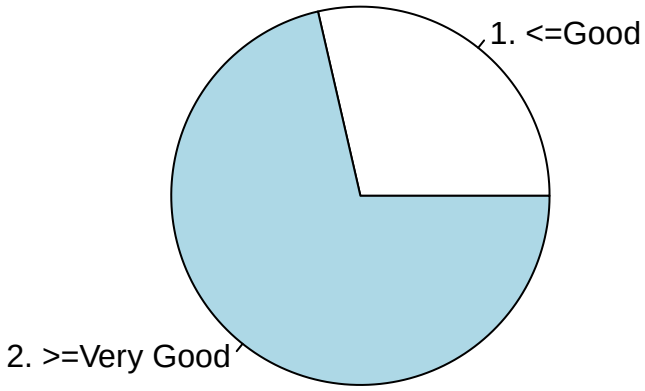
```
plot(x = dataset$health, main = "Distribució variable health",xlab = "Possibles valors", ylab = "Freqüè
```




```
table(dataset$health)
```

```
##
##      1. <=Good 2. >=Very Good
##            858           2142
```

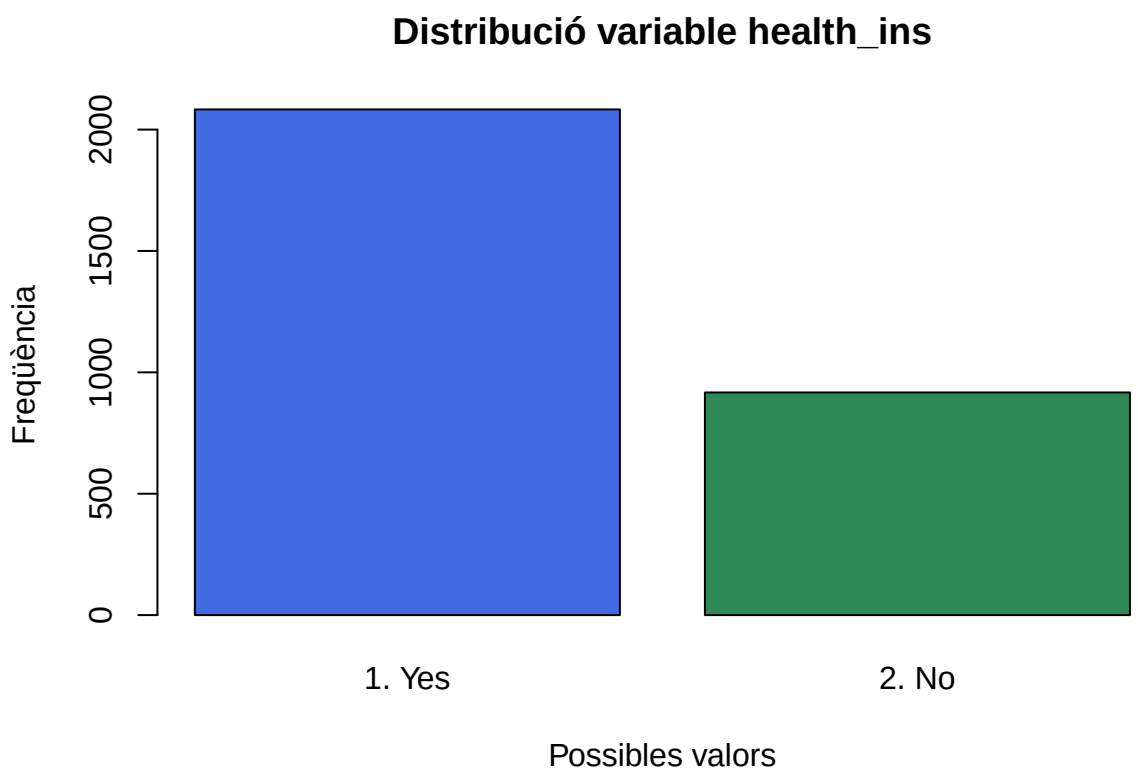
```
slices <- c(table(dataset$health))
lbls <- c("1. <=Good","2. >=Very Good")
pie(slices, labels = lbls, main="Distribucio poblacio segons health")
```



Variable de tipus factor que pot prendre dos valors i on predominen les persones amb un estat de salut molt bo (2142).

health_ins

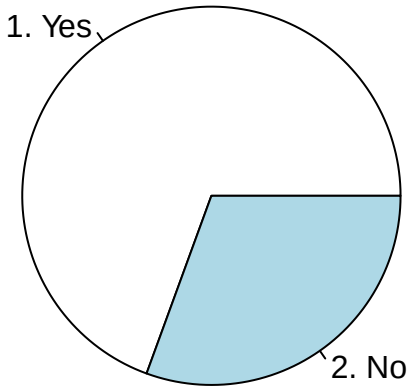
```
plot(x = dataset$health_ins, main = "Distribució variable health_ins",xlab = "Possibles valors", ylab =
```




```
table(dataset$health_ins)
```

```
##
## 1. Yes  2. No
##   2083    917
```

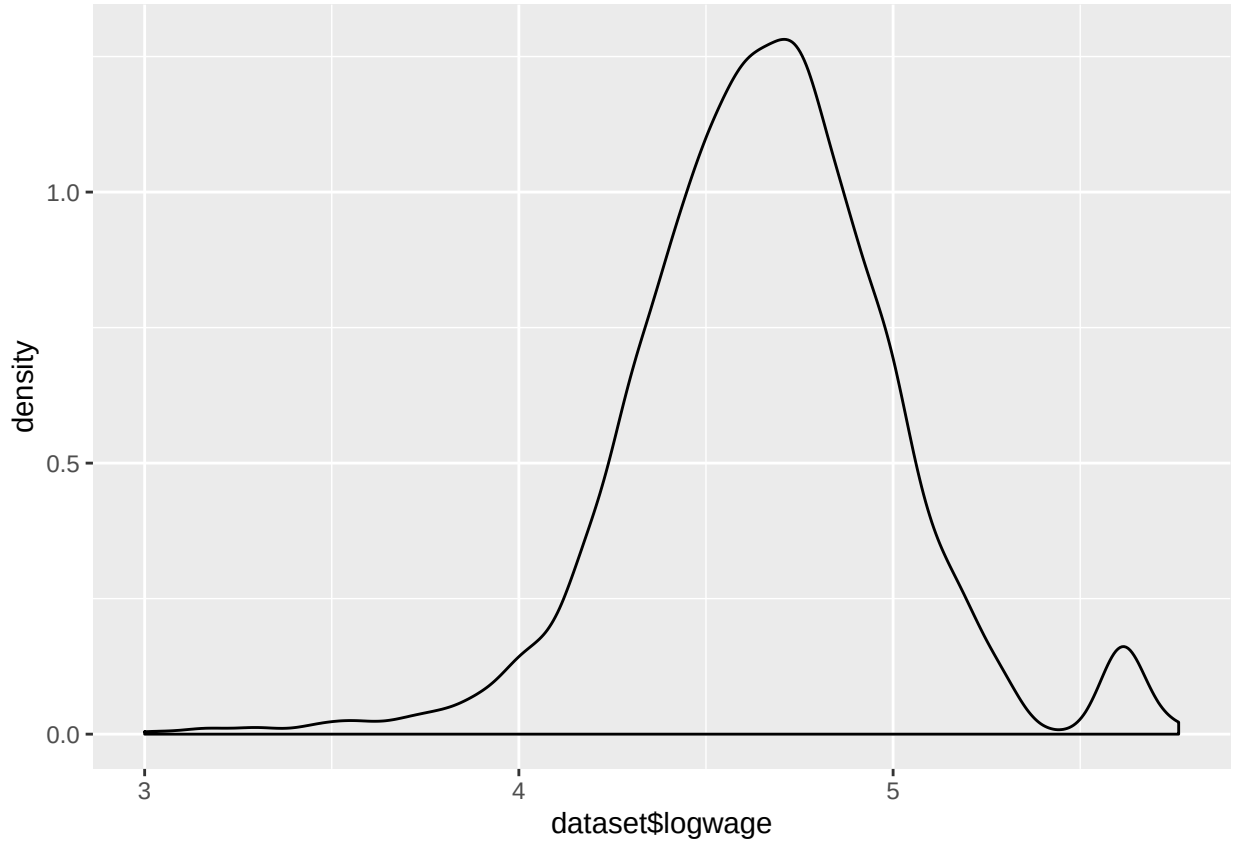
```
slices <- c(table(dataset$health_ins))
lbls <- c("1. Yes","2. No")
pie(slices, labels = lbls, main="Distribucio poblacio segons health_ins")
```



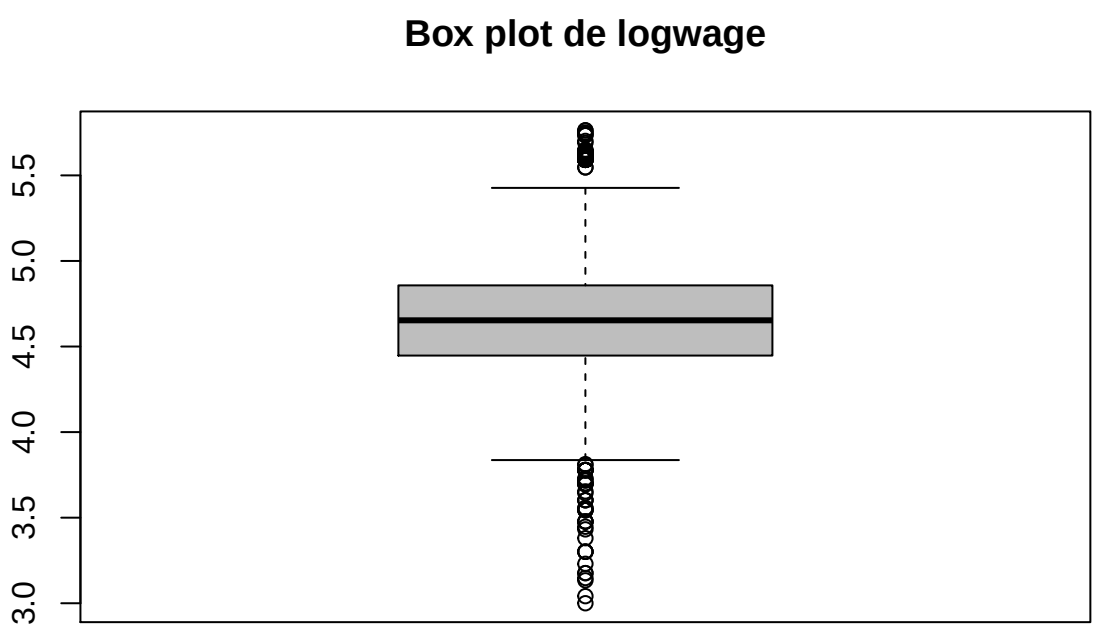
Variable de tipus factor que pren dos valors i on la majoria de treballadors presenten una assegurança mèdica privada. Concretament 2083 treballadors.

logwage

```
ggplot(mapping= aes(x=dataset$logwage))+ geom_density()
```




```
boxplot(dataset$logwage,main="Box plot de logwage", col="gray")
```



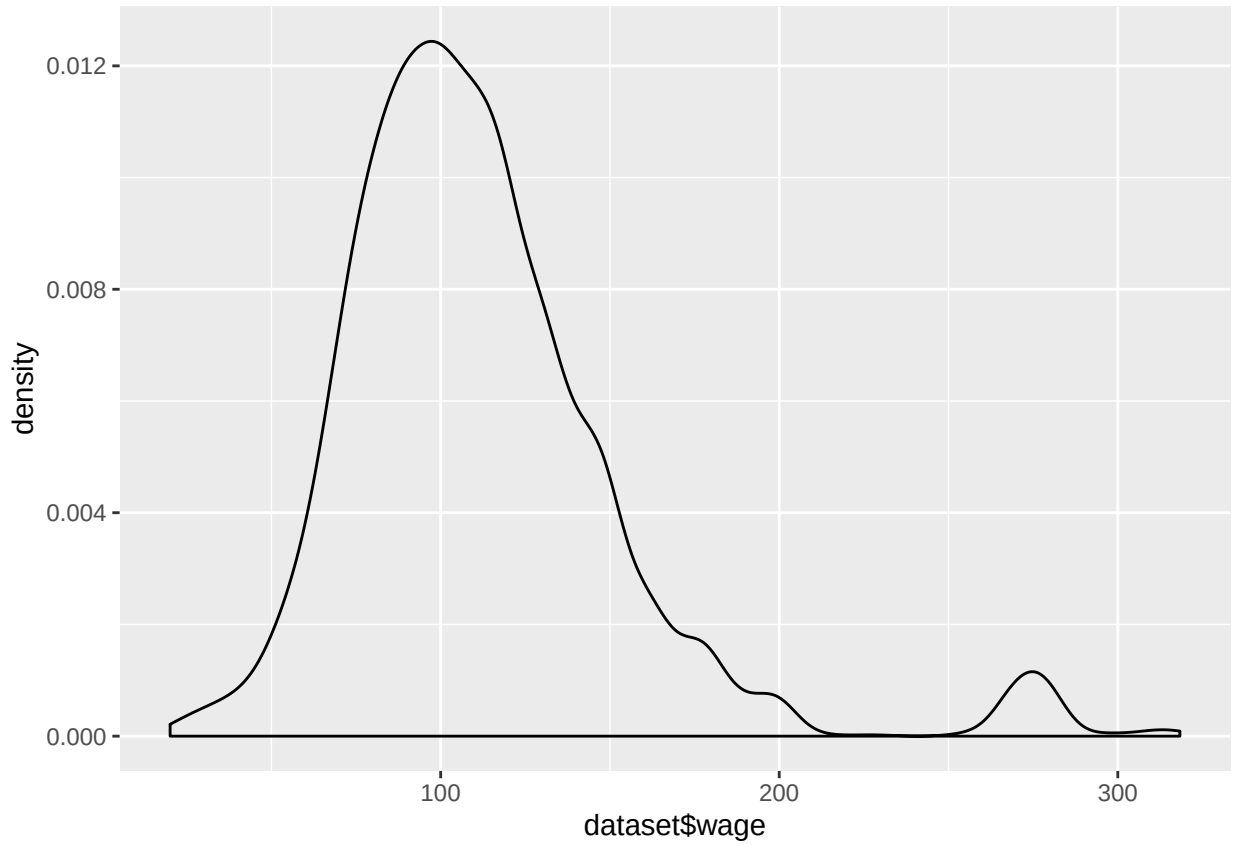

```
summary(dataset$logwage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   4.447   4.653   4.654   4.857   5.763
```

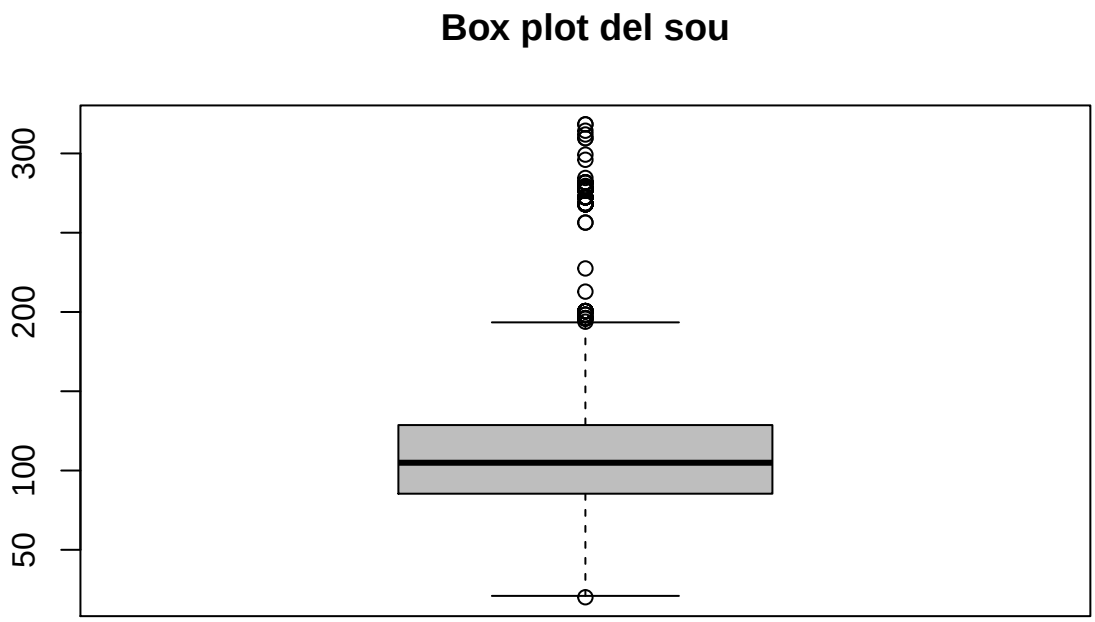
Variable de tipus numèrica continua on els valors es concentren al voltant de 4.5 i 5 (mitjana = 4.654). El valor màxim d'aquesta variable és 5.763 i el mínim 3. Es mostren també els sis valors resum d'una variable (mínim, màxim, Q1, mediana, mitjana i Q3) per tal de consolidar coneixement.

wage

```
ggplot(mapping= aes(x=dataset$wage))+ geom_density()
```




```
boxplot(dataset$wage,main="Box plot del sou", col="gray")
```




```
summary(dataset$wage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.09   85.38  104.92  111.70  128.68  318.34
```

Variable continua numèrica amb un sou mitjà de 111.70 ($/1000). El sou mínim recollit en el dataframe és 20.09 i el sou màxim 318.34. Es mostren també els sis valors resum d'una variable (mínim, màxim, Q1, mediana, mitjana i Q3) per tal de consolidar coneixement.

Selecció de grups d'interés a estudiar

A continuació seleccionem els grups del nostre conjunt de dades que hem considerat interessants per estudiar els sous en funció d'ells.

```
## Segons si es dediquen al món de la indústria o de la informació
industrial <- dataset[(dataset$jobclass == "1. Industrial"),]
informacio <- dataset[(dataset$jobclass == "2. Information"),]

## Segons si tenen assegurança mèdica privada
polissa <- dataset[(dataset$health_ins == "1. Yes"),]
no_polissa <- dataset[(dataset$health_ins == "2. No"),]

## Segons l'ètnia de la persona
blancs <- dataset[(dataset$race == "1. White"),]
```

```
negres <- dataset[(dataset$race == "2. Black"),]
asiatics <- dataset[(dataset$race == "3. Asian"),]
altres <-  dataset[(dataset$race == "4. Other"),]
```

Comprovació de la normalitat i homogeneïtat de la variància

Farem la prova de Shapiro wilk [2]. El test Shapiro Wilk es fa servir per contrastar la normalitat d'un conjunt de dades. Es planteja com a hipòtesi nul · la que una mostra x1, . . . , xn específicament d'una població normalment distribuïda. [3]

```
alpha = 0.05
col.names = colnames(dataset)
for (i in 1:ncol(dataset)) {
  if (i == 1) cat("Variables que no segueixen una distribució normal:\n")
  if (is.integer(dataset[,i]) | is.numeric(dataset[,i])) {
    p_val = shapiro.test(dataset[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
      # Format output
      if (i < ncol(dataset) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}
```

```
## Variables que no segueixen una distribució normal:
## year, age, logwagewage
```

Tanmateix, pel teorema del límit central assumim normalitat en tenir més de 30 mostres. El Teorema del límit central indica que la distribució de la suma estandarditzada de variables aleatòries independents que tenen una variància finita tendeix a una distribució normal estàndard quan el nombre de termes de la suma creix indefinidament. [4]

Ara és torn d'estudiar l'homogeneïtat de les variàncies. També conegut com test d'homoscedasticitat considera que la variància és constant entre els diferents grups d'una mostra, és a dir, tots els grups conformats anteriorment tenen la mateixa variància. Per comprovar-ho aplicarem el test de Fligner-Killen[5].

```
fligner.test(wage ~ jobclass, data = dataset)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  wage by jobclass
## Fligner-Killeen:med chi-squared = 19.166, df = 1, p-value =
## 1.198e-05
```

Aplicació de proves estadístiques per comparar els grups de dades.

Anàlisi de correlació

Començarem fent un anàlisi de correlació entre les diferents variables per poder determinar quina d'elles té més influència a l'hora de determinar el salari d'una persona. Per fer-ho, utilitzarem el coeficient de

correlació de Pearson, ja que hem assumim una distribució normal de tots els atributs pel teorema del límit central.

```
corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")
# Calcular el coeficiente de correlación para cada variable cuantitativa
# con respecto al campo "precio"
for (i in 1:(ncol(dataset) - 1)) {
  if (is.integer(dataset[,i]) | is.numeric(dataset[,i])) {
    spearman_test = cor.test(dataset[,i],dataset[,length(dataset)],method = "pearson")
    corr_coef = spearman_test$estimate
    p_val = spearman_test$p.value
    # Add row to matrix
    pair = matrix(ncol = 2, nrow = 1)
    pair[1][1] = corr_coef
    pair[2][1] = p_val
    corr_matrix <- rbind(corr_matrix, pair)
    rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(dataset)[i]
  }
}
print(corr_matrix)
```

```
##           estimate      p-value
## year    0.06554428 3.276778e-04
## age     0.19563720 2.900778e-27
## logwage 0.95068337 0.000000e+00
```

Podem identificar de les diferents variables numèriques, quines estan més relacionades amb el salari en funció de la seva proximitat als valors +-1. Observem que la variable numèrica més rellevant és *logwage*, i la que menys és *year*, el que significa que el registre de salari d'un treballador té més relació amb el salari de la persona que l'edat que té o que l'any en què es va recollir la mostra.

Interval de confiança de la variable *wage* segons ètnia

La segona prova estadística que realitzarem serà el càlcul dels intervals de confiança de la variable *wage* segons les diferents ètnies que es troben al dataset. D'aquesta manera veurem si segons l'ètnia els ingressos dels habitants són diferents i si existeix alguna desigualtat. Per fer aquesta prova utilitzarem els grups generats anteriorment on hem separat per la variable *race*.

Comencem calculant les mitjanes i la desviació estàndard de cada una de les mostres:

```
mitjana_blancs = mean(blancs$wage)
mitjana_negres = mean(negres$wage)
mitjana_asiatics = mean(asiatics$wage)
mitjana_altres = mean(altres$wage)
sd_blancs = sd(blancs$wage)
sd_negres = sd(negres$wage)
sd_asiatics = sd(asiatics$wage)
sd_altres = sd(altres$wage)

cat('La mitjana de cada grup és:\n
    - Blancs: ', mitjana_blancs ,'\n
    - Negres: ', mitjana_negres ,'\n
```

```
    - Asiatics: ', mitjana_asiatics ,'\n
    - Altres: ', mitjana_altres ,'\n'
    )
```

```
## La mitjana de cada grup és:
##
##     - Blancs:  112.5637
##
##     - Negres:  101.6012
##
##     - Asiatics:  120.2883
##
##     - Altres:  89.97333
```

```
cat('La desviació estàndard de cada grup és:\n
    - Blancs: ', sd_blancs ,'\n
    - Negres: ', sd_negres ,'\n
    - Asiatics: ', sd_asiatics ,'\n
    - Altres: ', sd_altres ,'\n'
    )
```

```
## La desviació estàndard de cada grup és:
##
##     - Blancs:  41.73383
##
##     - Negres:  37.16249
##
##     - Asiatics:  46.42251
##
##     - Altres:  29.15353
```

A partir d'aquí, calculem l'error de la mitjana de cadascuna de les mostres dividint la desviació estàndard entre l'arrel de la mida de la mostra:

```
stderr_blancs = sd_blancs/sqrt(dim(blancs)[1])
stderr_negres = sd_negres/sqrt(dim(negres)[1])
stderr_asiatics = sd_asiatics/sqrt(dim(asiatics)[1])
stderr_altres = sd_altres/sqrt(dim(altres)[1])

cat("L'error de la mitjana de cada grup és:\n
    - Blancs: ", sd_blancs ,"\n
    - Negres: ", sd_negres ,"\n
    - Asiatics: ", sd_asiatics ,"\n
    - Altres: ", sd_altres ,"\n"
    )
```

```
## L'error de la mitjana de cada grup és:
##
##     - Blancs:  41.73383
##
##     - Negres:  37.16249
##
```

```
##     - Asiatics:  46.42251
##
##     - Altres:  29.15353
```

Ara podem calcular el valor crític de cada grup. Aquest és el punt $t(alfa)/2$ tal que $P(Z >= t(alfa)/2) = (alfa)/2$ on t és una variable N(0,1).

```
punt_critic_blancs = qt(1-0.05/2, dim(blancs)[1]-1)
punt_critic_negres = qt(1-0.05/2, dim(negres)[1]-1)
punt_critic_asiatics = qt(1-0.05/2, dim(asiatics)[1]-1)
punt_critic_altres = qt(1-0.05/2, dim(altres)[1]-1)
cat("El valor crític de cada grup és:\n
    - Blancs: ", punt_critic_blancs ,"\n
    - Negres: ", punt_critic_negres ,"\n
    - Asiatics: ", punt_critic_asiatics ,"\n
    - Altres: ", punt_critic_altres ,"\n"
    )
```

```
## El valor crític de cada grup és:
##
##     - Blancs:  1.960921
##
##     - Negres:  1.968121
##
##     - Asiatics:  1.972595
##
##     - Altres:  2.028094
```

Arribats aquí estem en disposició de calcular els intervals de confiança:

```
cint_blancs <- mitjana_blancs + c(-punt_critic_blancs,punt_critic_blancs)*stderr_blancs
cint_negres <- mitjana_negres + c(-punt_critic_negres,punt_critic_negres)*stderr_negres
cint_asiatics <- mitjana_asiatics + c(-punt_critic_asiatics,punt_critic_asiatics)*stderr_asiatics
cint_altres <- mitjana_altres + c(-punt_critic_altres,punt_critic_altres)*stderr_altres
cat("'L'interval de confiança amb un 95% de la variable *wage* per cada grup és:\n
    - Blancs: ", cint_blancs ,"\n
    - Negres: ", cint_negres ,"\n
    - Asiatics: ", cint_asiatics ,"\n
    - Altres: ", cint_altres ,"\n"
    )
```

```
## 'L'interval de confiança amb un 95% de la variable *wage* per cada grup és:
##
##     - Blancs:  110.9203 114.207
##
##     - Negres:  97.32828 105.8741
##
##     - Asiatics:  113.6449 126.9317
##
##     - Altres:  80.25306 99.6936
```

Amb els valors obtinguts, es pot interpretar el resultat dels intervals de confiança calculats:

Aquests ens permeten afirmar que amb una confiança del 95% que el sou d'una persona que pertanyi a algun d'aquests grups es trobarà dins d'aquest interval, és a dir, si una persona és blanca amb un 95 de confiança podrem afirmar que el seu sou estarà entre 110.92k i 114.21k.

En altres paraules, en el 95% de les mostres de la mateixa mida que les que hem estudiat, el valor de la mitjana mostral fa que l'interval obtingut contingui el veritable valor de la mitjana de sou dels treballadors.

Veiem per tant com els asiàtics són l'ètnia amb un interval de la variable *wage* més alt, seguit dels blancs. A partir d'aquí, trobem les altres dues ètnies, però amb una diferència més gran respecte aquests. Per tant, podem dir que si existeix una desigualtat entre races produïda possiblement per molts factors com per exemple, la diferència del nivell d'estudis entre ètnies.

```
table(blancs$education)
```

```
##
##       1. < HS Grad         2. HS Grad    3. Some College
##                211                822                532
##    4. College Grad 5. Advanced Degree
##                576                339
```

```
table(asiatics$education)
```

```
##
##       1. < HS Grad         2. HS Grad    3. Some College
##                 15                 31                 18
##    4. College Grad 5. Advanced Degree
##                 66                 60
```

```
table(negres$education)
```

```
##
##       1. < HS Grad         2. HS Grad    3. Some College
##                 31                105                 92
##    4. College Grad 5. Advanced Degree
##                 40                 25
```

```
table(altres$education)
```

```
##
##       1. < HS Grad         2. HS Grad    3. Some College
##                 11                 13                  8
##    4. College Grad 5. Advanced Degree
##                  3                  2
```

Podem veure com en proporció, la presència de blancs i asiàtics en nivell d'estudis avançats és més alta que en la resta. Per altra banda, negres i la categoria que engloba *Others* mentre que la resta es concentren en educació secundària i educació superior.

Existeix una diferència en el sou segons si el treballador té contractada una polissa privada?

A continuació volem donar resposta la següent pregunta següent: Les persones amb una pòlissa privada tenen un sou més alt?

La prova estadística que realitzarem ara serà un contrast d'hipòtesi sobre dues mostres per determinar si el sou de les persones que tenen contractada una pòlissa mèdica privada és superior a les persones que no la tenen.

Primer utilitzar els dos subconjunts generats abans per començar la prova estadística.

```
n_row_polissa <- dim(polissa)[1]
n_row_no_polissa <- dim(no_polissa)[1]
cat("El nombre de treballadors amb pòlissa és: ", n_row_polissa , "\n")
```

```
## El nombre de treballadors amb pòlissa és:  2083
```

```
cat("El nombre de treballadors sense pòlissa és: ", n_row_no_polissa, "\n")
```

```
## El nombre de treballadors sense pòlissa és:  917
```

```
mean(polissa$wage)
```

```
## [1] 120.2383
```

```
mean(no_polissa$wage)
```

```
## [1] 92.3167
```

Dit això, la nostra hipòtesi nul · la serà que la mitjana del sou de les persones que tenen una pòlissa d'assegurança mèdica variable és la mateixa que la de les persones que no en tenen. Per contra, la hipòtesi alternativa serà que els treballadors amb una pòlissa d'assegurança mèdica variable tenen un sou mitjà 20 $ més alt que les persones sense pòlissa.

Per tant:

$Ho : muP = muNP$

$H1 : muP > muNP + 20$

On: P treballadors amb pòlissa

NP treballadors sense pòlissa

Aquestes dues mostres són independents perquè recullen dades sobre individus diferents. A més, simplement observant que la mida de les mostres és diferent podríem ja assegurar que són independents.

Com que desconeixem les variàncies poblacionals, i com que la mida dels dos datasets té una mida superior a 30 assumirem normalitat N pel teorema del límit central.

Podem dir que és un test paramètric perquè com hem dit els dos subsets segueixen una distribució estadística normal.

Finalment sabem que serà contrast unilateral perquè només es compara en una direcció, és a dir, si el sou mitjà és igual o 20 $ més alt en els treballadors amb pòlissa.

Concloíem, que com desconeixem les variàncies poblacionals, però les dues mostres són superior a 30 i no sabem si les dues mostres tenen la mateixa variància, assumim normalitat i suposem que es distribuiran aproximadament com una N(0,1).

Càlculs

Primer de tot calcularem manualment l'estadístic de contrast i el *p_value* i per comprovar que s'han calculat correctament farem servir la funció *t.test* que ens calcularà ..... (explicar què fa)

Primer, calcularem les mitjanes i les desviacions estàndards dels dos subsets:

```
mx_p <- mean(polissa$wage)
cat("Mitjana de sou treballadors amb pòlissa: ", mx_p , "\n")
```

```
## Mitjana de sou treballadors amb pòlissa:  120.2383
```

```
mx_np <- mean(no_polissa$wage)
cat("Mitjana de sou treballadors sense pòlissa: ", mx_np , "\n")
```

```
## Mitjana de sou treballadors sense pòlissa:  92.3167
```

```
std_p <- sd(polissa$wage)
cat("Desviació estàndard sou treballadors amb pòlissa: ", std_p , "\n")
```

```
## Desviació estàndard sou treballadors amb pòlissa:  41.23698
```

```
std_np <- sd(no_polissa$wage)
cat("Desviació estàndard sou treballadors sense pòlissa: ", std_np , "\n")
```

```
## Desviació estàndard sou treballadors sense pòlissa:  35.9719
```

Calculem amb això l'error estàndard. Aquest és:

$sqrt((std_p^2/nrow(polissa)) + (stdnp^2/nrow(nopolissa)))$

On: *std_p* és la desviació de la mostra de treballadors amb pòlissa

std_np és la desviació de la mostra de treballadors sense pòlissa contractada

n_row_polissa mida del subset de treballadors amb pòlissa.

n_row_no_polissa mida del subset de treballadors amb pòlissa.

```
std <- sqrt((std_p^2 / n_row_polissa) + (std_np^2 / n_row_no_polissa))
cat("L'error estàndard és:", std)
```

```
## L'error estàndard és: 1.492469
```

Estem en disposició de calcular l'estadístic de contrast:

```
z = (mx_p-mx_np-20)/(std) ## -20 PERQUE ES 20 EUROS MÉS ALTA
cat("L'estadístic de contrast és:", z)
```

```
## L'estadístic de contrast és: 5.307721
```

Podem calcular ja el *p-value*:

```
p_value = 2*(pnorm(-abs(z)))
cat("El p-value és:", p_value)
```

```
## El p-value és: 1.110042e-07
```

El valor crític:

```
qnorm(1-0.95)
```

```
## [1] -1.644854
```

Un cop tenim els valors calculats, podem rebutjar la hipòtesi nul · la, és a dir, podem dir amb un nivell de confiança del 95%, que la mitjana dels sous dels treballadors amb pòlissa no és la mateixa que el dels treballadors sense pòlissa contractada.

Per comprovar-ho podem fixar-nos en el p-valor calculat, veiem que aquest és inferior al nivell de significança que estàvem buscant (5%).

$1.110042e-07 << 0.05$

Concloem per tant, que cal rebutjar la hipòtesi nul · la.

Hipòtesi alternativa

Ara, ens centrarem en la hipòtesi alternativa, calcularem l'interval amb un 95% de confiança de la diferència de mitjanes. Per fer-ho, calculem la variància dels subsets (*s pulled*).

```
df =  dim(dataset)[1] -1
qt = qt(p = 0.05, df = df, lower.tail = FALSE)

s_pulled = ((n_row_polissa-1)*sd(polissa$wage) + (n_row_no_polissa-1)*sd(no_polissa$wage))/(n_row_polis
err <- qt* sqrt(s_pulled/(n_row_polissa + n_row_no_polissa))

interval <- mx_p - mx_np + c(-err,err)
expected<- mx_p - mx_np


cat("L'interval amb un 95% de confiança de la diferència de mitjanes és: ",interval, "\n")
```

```
## L'interval amb un 95% de confiança de la diferència de mitjanes és:  27.73251 28.11072
```

Veiem com el valor esperat cau dintre de l'interval de confiança. Per tant, amb un 95% de confiança podem afirmar que el salari mitjà dels treballadors amb pòlissa privada és major que el dels treballadors que no tenen pòlissa contractada. A més, el nostre objectiu era confirmar que era 20 $ major, i podem assegurar-ho perquè tot l'interval està per sobre dels 20 de diferència.

Per comprovar que tot ha anat bé:

Comprovació t.test

```
t.test(polissa$wage, no_polissa$wage, alternative = "greater", mu=20)
```

```
##
##  Welch Two Sample t-test
##
## data:  polissa$wage and no_polissa$wage
## t = 5.3077, df = 1989.5, p-value = 6.169e-08
## alternative hypothesis: true difference in means is greater than 20
## 95 percent confidence interval:
##  25.46557      Inf
## sample estimates:
## mean of x mean of y
##  120.2383   92.3167
```

Comprovem que amb el t.test obtenim el mateix estadístic de contrast i la p-value és menor.

Regressió logística

Farem un model predictiu basat en la regressió logística per predir la probabilitat de tenir un salari superior a la mitjana en funció de les variables: *health_ins*, *jobclass* i *age*. Aquesta regresió logística ens permetrà predir el resultat d'una variable categòrica en funció de les variables independents anteriors. Primerament, cal crear una nova variable al dataframe que reculli un valor 0 quan el salari sigui inferior a la mitjana mostral i 1 quan el salari sigui superior a la mitjana mostral.

```
mean <- mean(dataset$wage)
dataset$sou_factoritzat[dataset$wage < mean] <-  0
dataset$sou_factoritzat[dataset$wage >= mean] <-  1
```

La variable generada:

```
str(dataset$sou_factoritzat)
```

```
##  num [1:3000] 0 0 1 1 0 1 1 1 1 1 ...
```

```
table(dataset$sou_factoritzat)
```

```
##
##    0    1
## 1724 1276
```

Un cop disposem de la nova variable, podem estimar un model de regressió logística on la variable dependent sigui *sou_factoritzat* i les variables explicatives siguin les demanades: *health_ins*, *jobclass* i *age*.

```
model_reg_log = glm(formula = sou_factoritzat~ health_ins + jobclass + age, data = dataset, family = bi
summary(model_reg_log)
```

```
##
## Call:
## glm(formula = sou_factoritzat ~ health_ins + jobclass + age,
##     family = binomial, data = dataset)
##
## Deviance Residuals:
```

```
##    Min      1Q  Median      3Q     Max
## -1.752  -1.042  -0.586   1.104   2.062
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.373068   0.163067  -8.420  < 2e-16 ***
## health_ins2. No        -1.224173   0.093342 -13.115  < 2e-16 ***
## jobclass2. Information  0.586006   0.078709   7.445 9.68e-14 ***
## age                     0.025984   0.003509   7.405 1.31e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4091.7  on 2999  degrees of freedom
## Residual deviance: 3726.2  on 2996  degrees of freedom
## AIC: 3734.2
##
## Number of Fisher Scoring iterations: 4
```

Obtenim un $AIC = 3734.2$. Si ens fixem en els regressors que expliquen el model, tots tenen una gran influència perquè el seu p-valor de contrast individual és menor del 5%. En especial, el que sembla tenir una influència més alta és el fet de tenir una pòlissa variable o no, però podem afirmar que tots són significatius.

Per tant, podem afirmar que els tres regressors expliquen bé la variable *wage*.

```
model_reg_log = glm(formula = sou_factoritzat~ health_ins + jobclass + age + education, data = dataset,
summary(model_reg_log)
```

```
##
## Call:
## glm(formula = sou_factoritzat ~ health_ins + jobclass + age +
##     education, family = binomial, data = dataset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2149  -0.8726  -0.4314   0.8940   2.5503
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -2.788611   0.264697 -10.535  < 2e-16 ***
## health_ins2. No            -1.108971   0.102322 -10.838  < 2e-16 ***
## jobclass2. Information      0.105917   0.088599   1.195  0.23191
## age                         0.028543   0.003862   7.390 1.47e-13 ***
## education2. HS Grad         0.647155   0.206820   3.129  0.00175 **
## education3. Some College    1.363390   0.210033   6.491 8.51e-11 ***
## education4. College Grad    2.214955   0.209626  10.566  < 2e-16 ***
## education5. Advanced Degree 3.190188   0.233075  13.687  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4091.7  on 2999  degrees of freedom
```

```
## Residual deviance: 3242.6  on 2992  degrees of freedom
## AIC: 3258.6
##
## Number of Fisher Scoring iterations: 4
```

A continuació farem una predicció de la probabilitat de superar el salari mitjà un treballador de 42 anys, amb pòlissa mèdica amb formació de graduat i exercint en el món de la informació:

Per fer-ho, ens basarem en l'últim model generat amb totes les variables i utilitzarem la funció *predict*.

Comencem per crear el dataframe:

```
treballador <- data.frame(health_ins = "1. Yes",
                      jobclass = "2. Information",
                      age = 42,
                      education = "4. College Grad"
                      )
```

Executem la funció *predict* amb el parámetre *probability* a *TRUE* per tal que ens retorni la probabilitat de què el treballador tingui un sou superior a la mitjana:

```
predict(object = model_reg_log, newdata = treballador ,type="response", probability = TRUE)
```

```
##         1
## 0.6750432
```

Observem que hi ha una probabilitat del 67.50 de què un treballador de 42 anys tingui un sou superior o igual a la mitjana, sent graduat universitari, amb pòlissa privada i dedicant-se a l'àmbit de la informació. Això ho deduïm perquè en els models de regressió logística la variable dependent ha de prendre valors qualitatius, és a dir, ha de ser categòrica. En aquest cas (0,1). Com que veiem que el resultat retornat és 0.6750, afirmem que la possibilitat que *sou_factoritzat* sigui 1 és de 67.50%.

Si ara assumim que el treballador es dedica a l'àmbit industrial, només caldrà canviar el valor de la variable *jobclass* i tornar a aplicar el *predict*:

```
treballador <- data.frame(health_ins = "1. Yes",
                      jobclass = "1. Industrial",
                      age = 42,
                      education = "4. College Grad"
                      )

predict(object = model_reg_log, newdata = treballador ,type="response", probability = TRUE)
```

```
##         1
## 0.6513929
```

En aquest cas, veiem que la probabilitat ha disminuït una mica 65.14, és a dir, que un treballador amb les mateixes característiques però que es dediqui a la indústria té un 65.14% de probabilitats de tenir un sou per sobre de la mitjana o igual a aquesta.

Per tant, podem concloure que és més probable que si dues persones tenen 42 anys, pòlissa privada i estudis universitaris, si aquesta es dedica al món de la informació, rebi un sou més alt o igual a la mitjana que una que es dediqui al món de la indústria.

Ara, tornem a generar el model de regressió, però aquest cop no volem predir si els treballadors tenen un sou per sobre de la mitjana, sinó que volem predir quin serà el seu sou. D'aquesta manera confirmarem el que hem vist amb l'anterior model: quin dels dos treballadors tindrà un sou més alt.

Farem servir les mateixes variables i l'única diferència és que ara la variable d'estudi torna a ser *wage*.

```
model_reg_log = glm(formula = wage~ health_ins + jobclass + age + education, data = dataset)
summary(model_reg_log)
```

```
##
## Call:
## glm(formula = wage ~ health_ins + jobclass + age + education,
##     data = dataset)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -101.790   -19.551    -3.764    13.740   214.314
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   73.75238    3.35288  21.997  < 2e-16 ***
## health_ins2. No              -18.36717    1.43747 -12.777  < 2e-16 ***
## jobclass2. Information         3.01573    1.35001   2.234 0.025566 *
## age                            0.46281    0.05622   8.232 2.73e-16 ***
## education2. HS Grad            8.24247    2.42387   3.401 0.000681 ***
## education3. Some College      18.82182    2.57420   7.312 3.37e-13 ***
## education4. College Grad      33.25266    2.58361  12.871  < 2e-16 ***
## education5. Advanced Degree   57.27546    2.83455  20.206  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1221.532)
##
##     Null deviance: 5222086  on 2999  degrees of freedom
## Residual deviance: 3654822  on 2992  degrees of freedom
## AIC: 29847
##
## Number of Fisher Scoring iterations: 2
```

Un cop generat el nou model estem en disposició de tornar a executar la funció *predict* per obtenir els sous dels treballadors:

```
predict(object = model_reg_log, newdata = treballador)
```

```
##        1
## 126.4433
```

```
treballador <- data.frame(health_ins = "1. Yes",
                     jobclass = "2. Information",
                     age = 42,
                     education = "4. College Grad"
                     )
predict(object = model_reg_log, newdata = treballador)
```

```
##       1
## 129.459
```

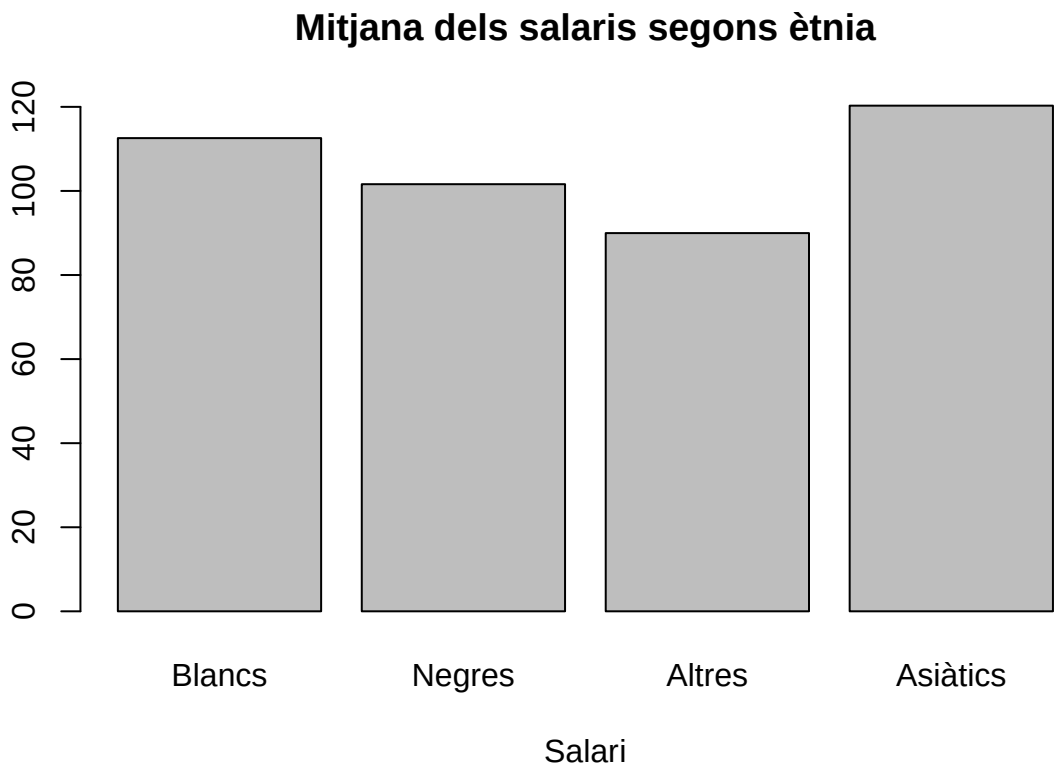
Com era d'esperar veiem com el treballador dedicat al sector de la informació tindria un sou més alt.

Representació de resultats

Hi ha diferents representacions que ens poden mostrar la diferència salarial segons els atributs que han estat objecte d'estudi en les proves estadístiques. Un dels més significatius és la diferència de sou segons l'ètnia.

```
options(scipen=5)
blancs_wage = mean(blancs$wage)
negres_wage = mean(negres$wage)
altres_wage = mean(altres$wage)
asiatics_wage = mean(asiatics$wage)

counts <- c(blancs_wage, negres_wage, altres_wage, asiatics_wage)
barplot(counts, names=c("Blancs", "Negres", "Altres", "Asiàtics"), main="Mitjana dels salaris segons èt
    xlab="Salari")
```

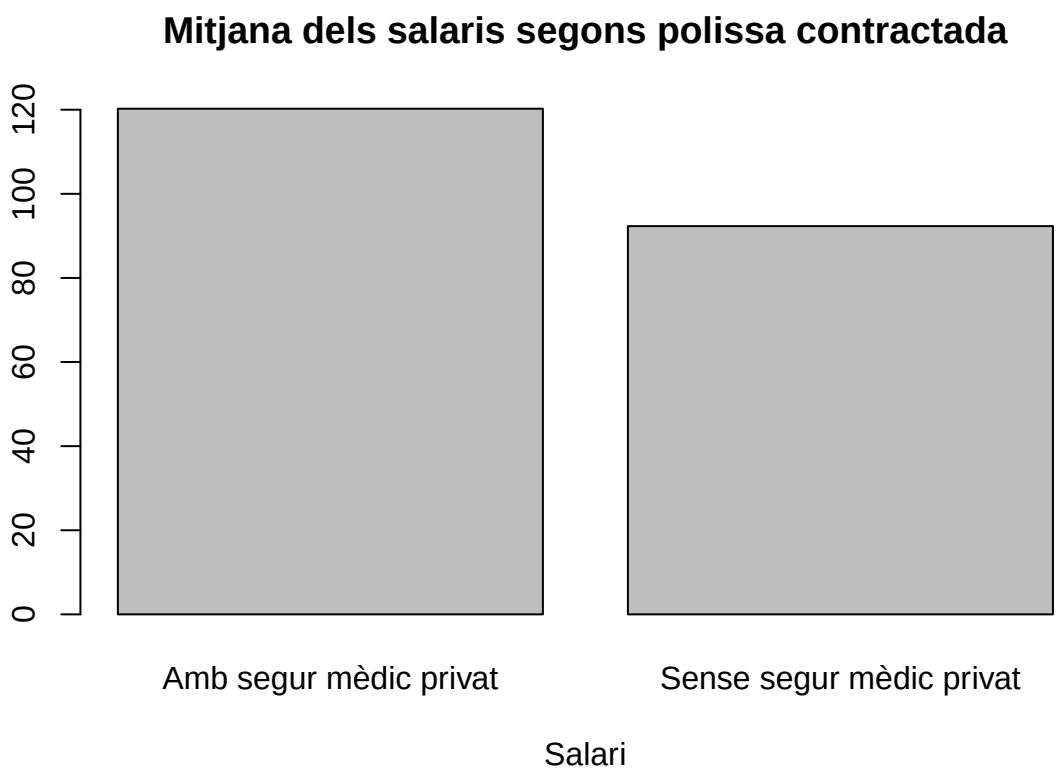


Com ja havíem pogut veure en les proves anteriors, els blancs i els asiàtics són els que més cobren.

Visualitzem ara la diferència de sou segons si tenen pòlissa contractada.

```
polissa_wage = mean(polissa$wage)
no_polissa_Wage = mean(no_polissa$wage)
```

```
counts <- c(polissa_wage, no_polissa_Wage)
barplot(counts, names=c("Amb segur mèdic privat", "Sense segur mèdic privat"), main="Mitjana dels salari
    xlab="Salari")
```



Un altre cop com havíem vist a les proves estadístiques, les persones amb un segur mèdic privat tenen una mitjana de sou més gran.

Conclusions

Com hem pogut veure s'han realitzat diferents proves estadístiques sobre el dataset. Hem començat fent un estudi dels valors nuls i extrems per determinar si havíem de fer algun tractament a les tuples del dataset. Hem vist que no feia falta, però el que sí que hem fet, ha estat modificar el nombre d'atributs, ja que hem vist que existia un atribut que no aportava informació rellevant sobre les dades i sobre l'objectiu que ens havíem establert.

De l'anàlisi de correlació hem vist que la variable numérica més rellevant és *logwage*, i la que menys és *year*, el que significa que el registre de salari d'un treballador té més relació amb el salari de la persona que l'edat que té o que l'any en què es va recollir la mostra.

Seguidament hem trobat l'interval de confiança de la variable *wage* segons l'ètnia. Hem observat que els asiàtics i els blancs són els que cobren més. Analitzant el perquè, segurament influeixen molts factors externs que no es mostren al dataset com podria ser l'entorn familiar, l'accés als estudis, etc., però un tòpic que sembla rellevant és el nivell d'estudis. Aquí s'ha observat com la majoria de registres que es troben en el rang de nivells d'estudis més alts pertanyen a l'ètnia blanca o asiàtica.

A continuació, hem realitzat també un contrast d'hipòtesi, on hem pogut concloure que amb un nivell de confiança del 95%, la mitjana dels sous dels treballadors amb pòlissa no és la mateixa que el dels treballadors sense pòlissa contractada, sent la gent amb pòliça la que ingressa més diners.

Finalment, hem generat models de regressió logística en funció de diverses variables explicatives: *health_ins*, *jobclass* i *age* per estudiar el sou dels treballadors. En aquest cas, hem vist que donats dos treballadors amb les mateixes característiques, però dedicats a sectors diferents (*information* o *îndustrial*), el primer tindrà un sou més alt.

Bibliografia

[1] Rdocumentation.org. 2020. Wage Function | R Documentation. [online] Available at: https://www.rdocumentation.org/packages/ISLR/versions/1.2/topics/Wage [Accessed 7 June 2020].

[2] Rdocumentation.org. 2020. Shapiro.Test Function | R Documentation. [online] Available at: https://www.rdocumentation.org/packages/stats/versions/3.6.1/topics/shapiro.test [Accessed 7 June 2020].

[3] Sthda.com. 2020. Normality Test In R - Easy Guides - Wiki - STHDA. [online] Available at: http://www.sthda.com/english/wiki/normality-test-in-r [Accessed 7 June 2020].

[4] En.wikipedia.org. 2020. Central Limit Theorem. [online] Available at: https://en.wikipedia.org/wiki/Central_limit_theorem [Accessed 7 June 2020].

[5] Rdocumentation.org. 2020. Fligner.Test Function | R Documentation. [online] Available at: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/fligner.test [Accessed 7 June 2020].