Pere Amat Riera - Pràctica 1: Informe EDA con R Studio ( Exploratory Data Analysis



# ÍNDEX

[**1. Càrrega de dades:**](#_7xqpc6hznr6d) **2**

[**2. Exploració inicial:**](#_m8ir49f0ftuk) **3**

[**3. Visualització univariada:**](#_buwll4cfb2lt) **3**

[Histograma:](#_yqr927f8vpn7) 4

[Grafico de barras](#_durwgc8m3u2z) 4

[BOXPLOT](#_rm9abifj1jvz) 5

[**4.Visualització bivariada:**](#_vfljpsu955cv) **6**

[GRAFICO DE DISPERSIÓ](#_70ov1nzauo8n) 6

[**5. Anàlisi de correlació:**](#_dghs5j6ea1k9) **7**

[**6. Gestió de valors perduts:**](#_us4pqbscgm9a) **9**

[**7. Anàlisi addicional:**](#_yqm88qja6x4d) **9**

# 

## 

## 

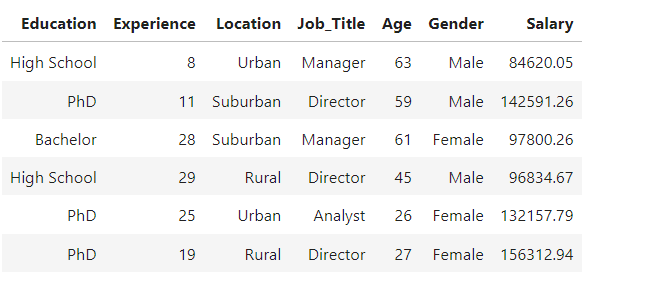
## 

## 1. Càrrega de dades:

En aquest primer pas, carregarem les dades utilitzant la funció read.csv(), que s’utilitza per llegir fitxers en format CSV i convertir-los en un objecte de tipus data frame

**df <- read.csv("C:/Users/pere amat/CE/PIA/PRACTICA1/salary\_prediction\_data.csv")**

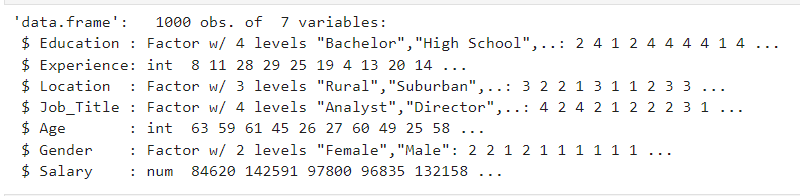
Per tenir una idea inicial de com són les dades carregades, utilitzem la funció **head()**, que ens mostra les primeres 6 files del data frame per defecte.



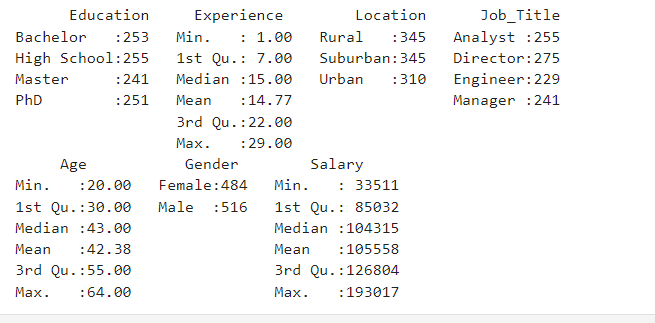
## 2. Exploració inicial:

En aquest pas, farem una exploració inicial del conjunt de dades per comprendre millor la seva estructura i el contingut.

Per veure l'estructura del conjunt de dades, podem utilitzar la funció str(). Aquesta funció ens mostrarà informació sobre el nombre de files i columnes, els noms de les columnes, així com els tipus de dades de cada columna.



Per obtenir una visió general dels valors numèrics del conjunt de dades, podem fer servir la **funció summary().**



## 3. Visualització univariada:

En aquest pas, ens centrarem en la visualització univariada per analitzar la distribució de les variables de manera individual. Crearem gràfics com histogrames, diagrames de barres i diagrames de caixa per identificar patrons, tendències i outliers

### Histograma:

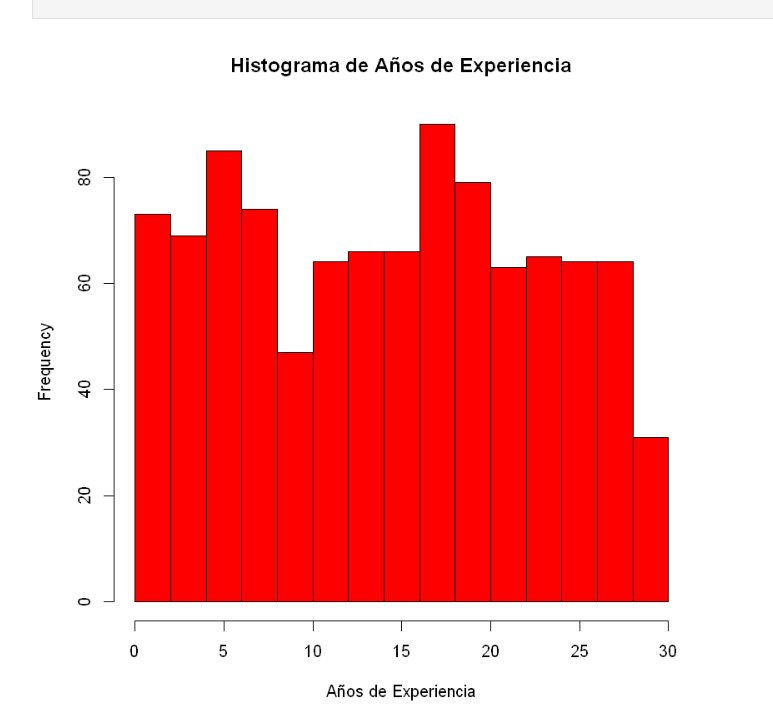
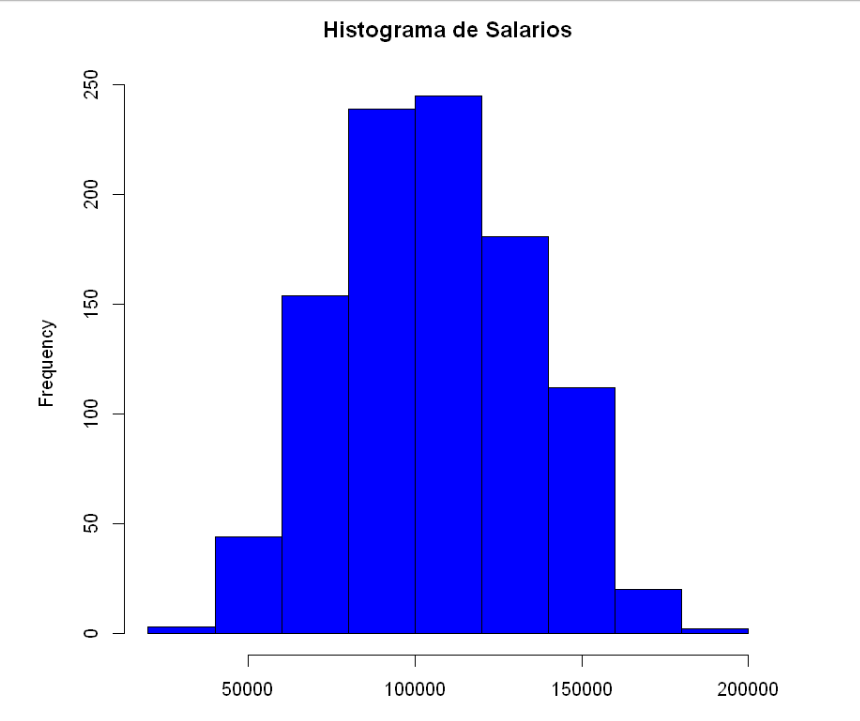
En aquest cas, utilitzem histogrames per analitzar la distribució de variables clau com els salaris, els anys d'experiència i l'edat dels individus del nostre conjunt de dades.

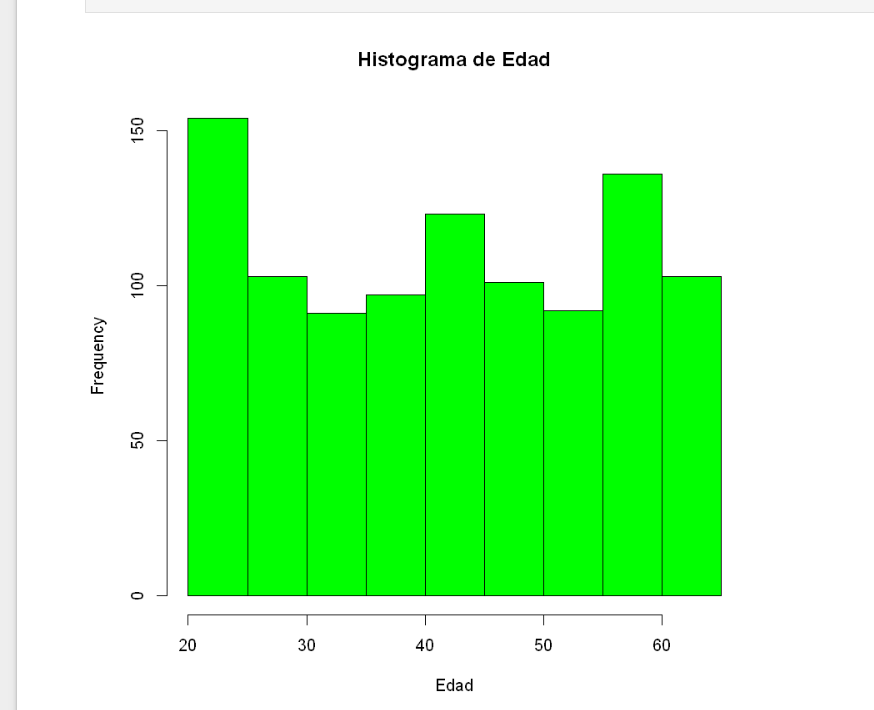
**hist(df$Salary, main="Histograma de Salarios",**

**xlab="Salarios", col="blue", breaks=10)**

Els histogrames mostren que els salaris es concentren entre 75.000 i 125.000 euros, amb un pic al voltant dels 100.000. Els anys d'experiència estan més dispersos, però hi ha concentracions entre 5 i 10 anys i entre 15 i 20 anys.

Pel que fa a l'edat, es distribueix de manera uniforme, amb lleugers pics als 20-30 i 50-60 anys.





### Gràfic de barras

**barplot(table(df$Education),**

**main="Frecuencia de Nivel Educativo",**

**xlab="Nivel Educativo", ylab="Frecuencia",**

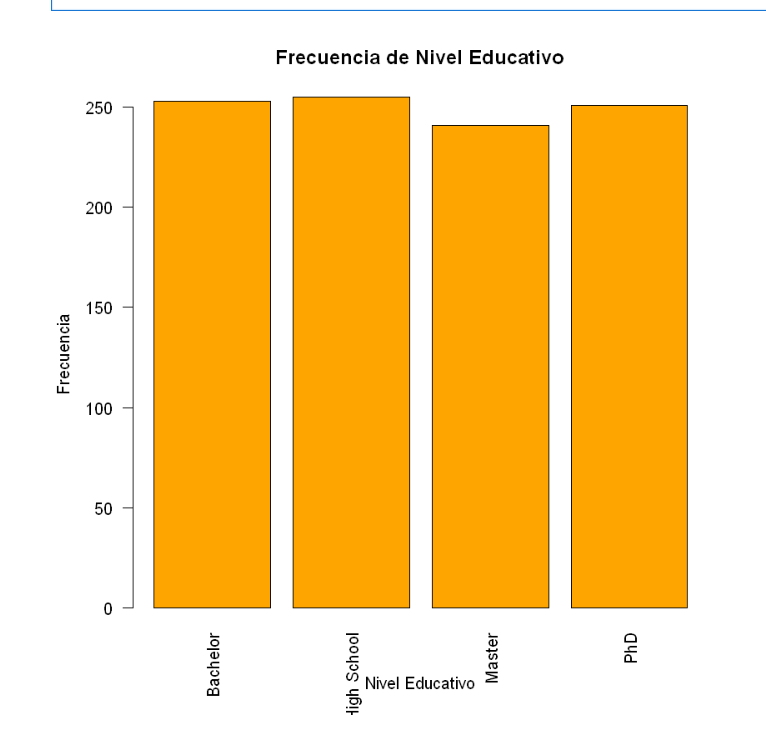
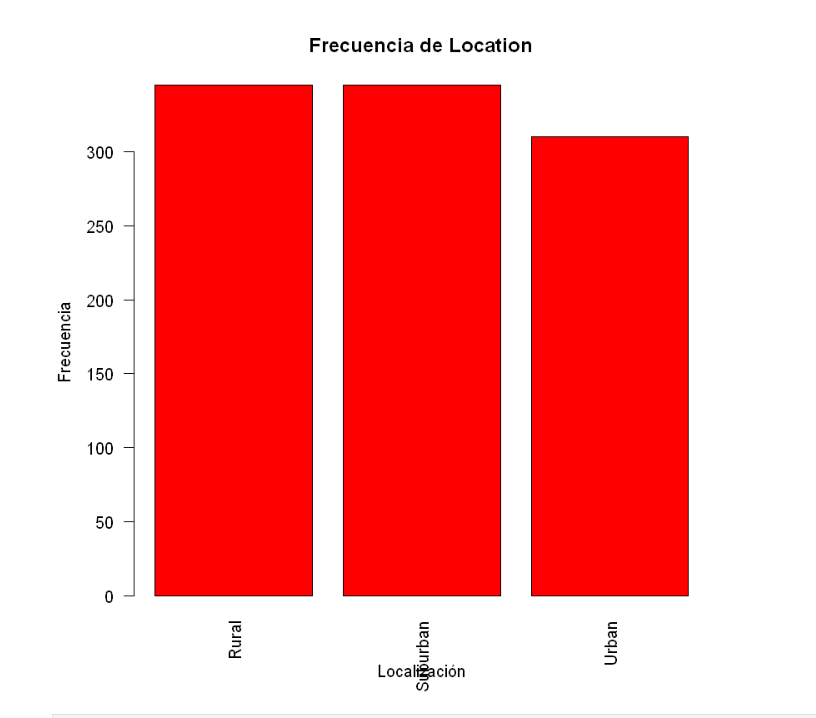
**col="orange", las=2)**

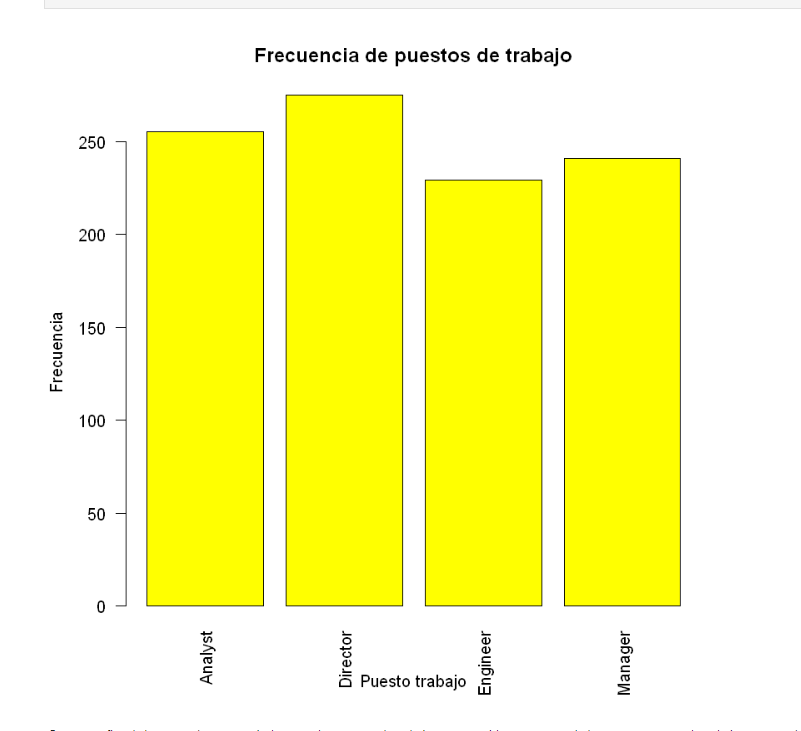
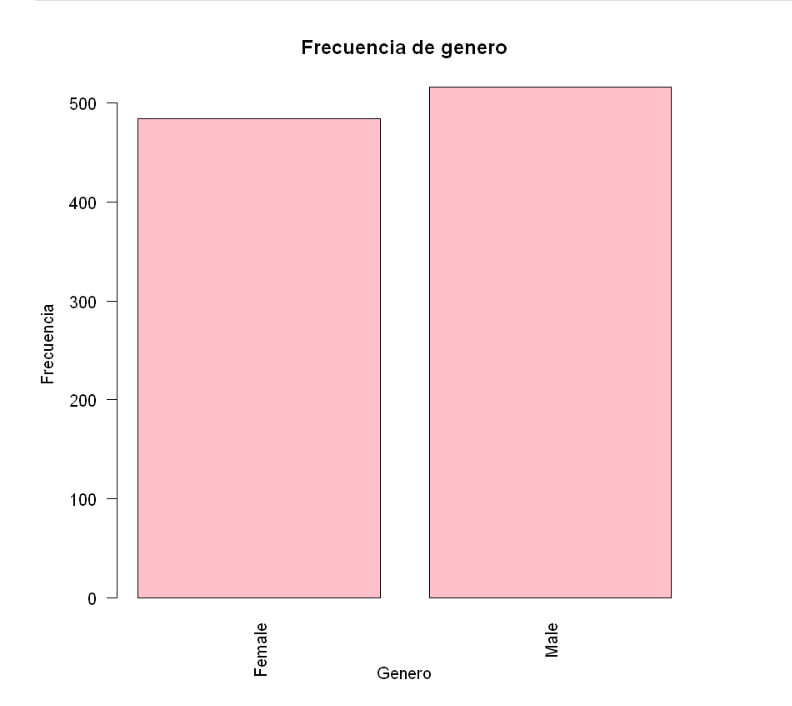
Els gràfics mostren una distribució bastant uniforme en el nivell educatiu, amb valors més elevats de persones amb títol de grau i PhD.

La ubicació es reparteix de manera equilibrada entre rural, suburbana i urbana.

Els càrrecs de treball estan també força repartits, però els directors són el grup més nombrós. Pel que fa al gènere, la distribució és pràcticament igual entre homes i dones.

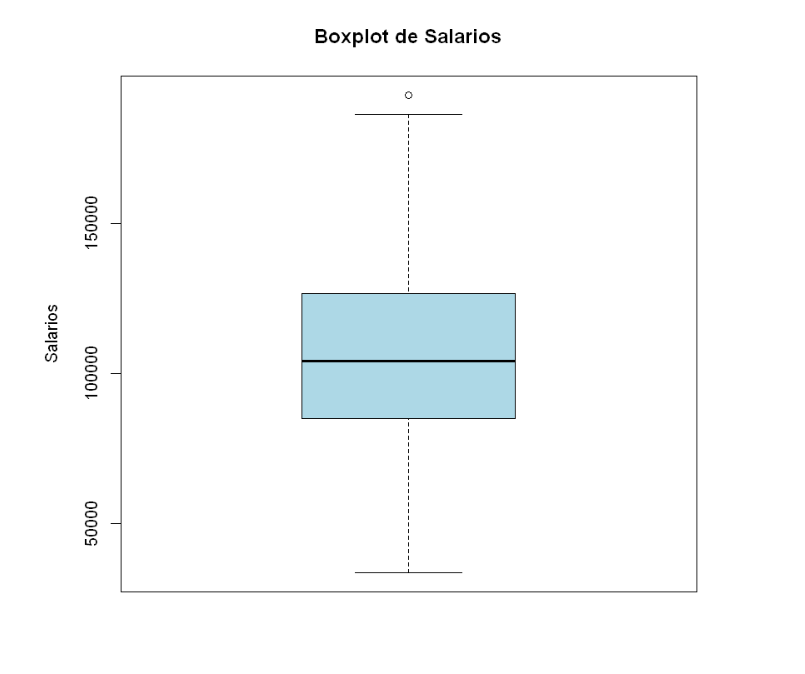
En general, les dades reflecteixen una mostra equilibrada tant en educació, ubicació, posicions laborals i gènere.





### BOXPLOT

El boxplot dels salaris mostra que la mediana es troba al voltant dels 100.000 euros. Els salaris es distribueixen entre uns 50.000 i 150.000 euros, amb algunes dades extremes (outliers).



## 

## 

## 

## 

## 

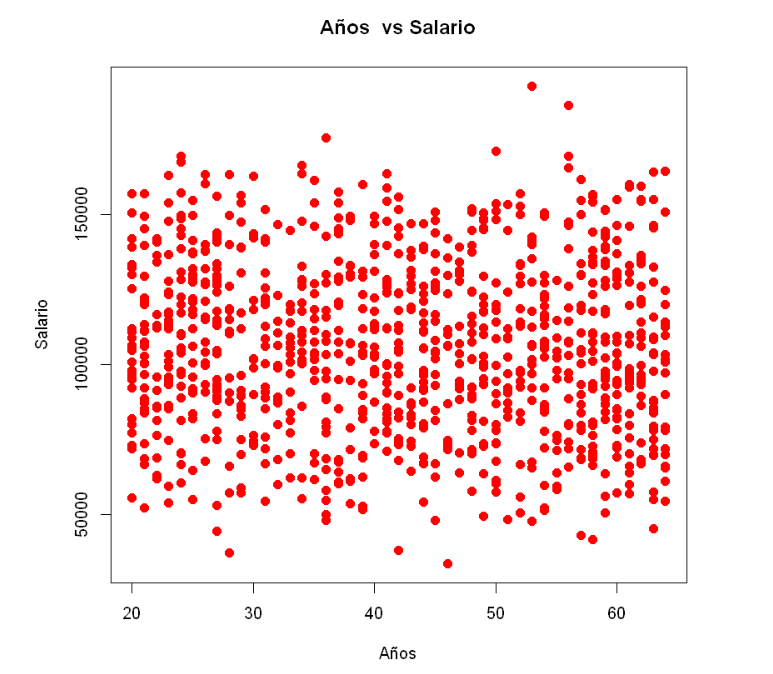
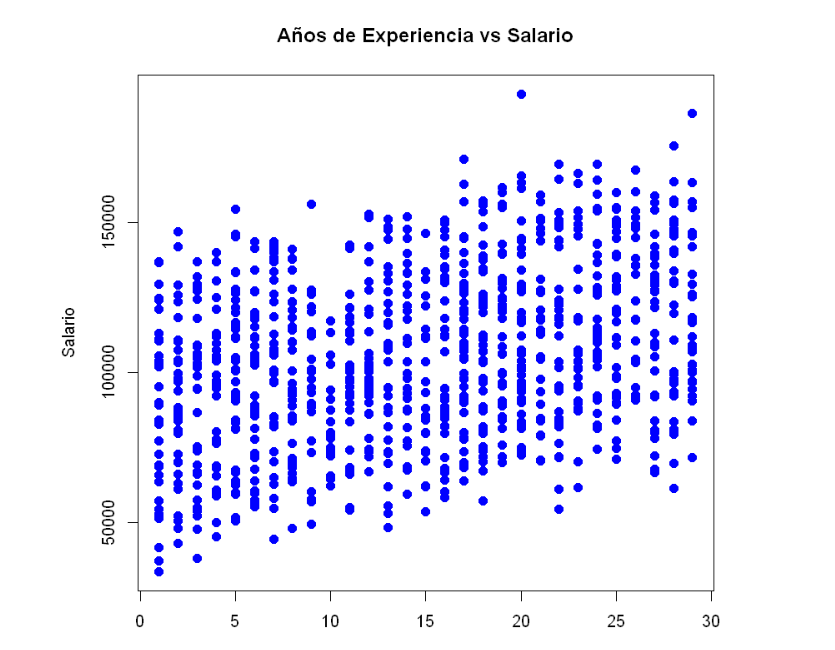
## 

## 4.Visualització bivariada:

### GRÀFIC DE DISPERSIÓ

Els gràfics mostren dues relacions entre salaris i variables com els anys d'experiència i l'edat. En el primer gràfic, hi ha una tendència positiva: a mesura que els anys d'experiència augmenten, els salaris.

En canvi, el segon gràfic no mostra una correlació clara entre l'edat i els salaris, amb una distribució més uniforme dels salaris a través de les diferents edats.



## 5. Anàlisi de correlació:

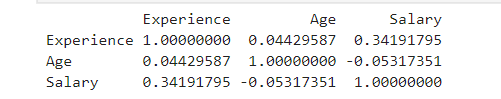
La matriu de correlació ens ajudarà a examinar les relacions lineals entre diverses variables d'un conjunt de dades.

**correlation\_matrix <- cor(df[, sapply(df, is.numeric)], use = "complete.obs")**

**print(correlation\_matrix)**

La correlació entre l'experiència i el salari és moderada (0.34), indicant que més experiència sovint significa salaris més alts. La relació entre experiència i edat és molt dèbil (0.04), i l'edat té una correlació negativa i dèbil amb el salari (-0.05).

Com es pot veure, l'experiència influeix en el salari, mentre que l'edat no sembla afectar-lo, cosa que segueix una logica amb el gràfic de dispersió observat anteriorment.



En l'exemple, les correlacions ens han ajudat a identificar quins valors tenen més importància. No obstant això, falten dades de variables com "educació", "ubicació" i "gènere", que són de tipus text, i m'agradaria incloure-les en l'anàlisi. Per això, he decidit convertir aquestes variables categòriques en variables numèriques.

**# Convertir la columna "Education" a numérico**

**df$Education <- as.numeric(factor(df$Education,**

**levels = unique(df$Education),**

**labels = 0:(length(unique(df$Education)) - 1)))**

**# Convertir la columna "Location" a numérico**

**df$Location <- as.numeric(factor(df$Location,**

**levels = unique(df$Location),**

**labels = 0:(length(unique(df$Location)) - 1)))**

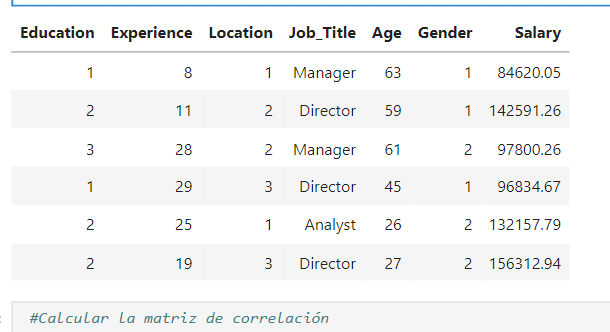
**# Convertir la columna "Gender" a numérico**

**df$Gender <- as.numeric(factor(df$Gender,**

**levels = unique(df$Gender),**

**labels = 0:(length(unique(df$Gender)) - 1)))**

**# Verificar la conversió**



**#Calcular la matriz de correlación**

**correlation\_matrix <- cor(df[, sapply(df, is.numeric)], use = "complete.obs")**

**# Mostrar la matriz de correlación**

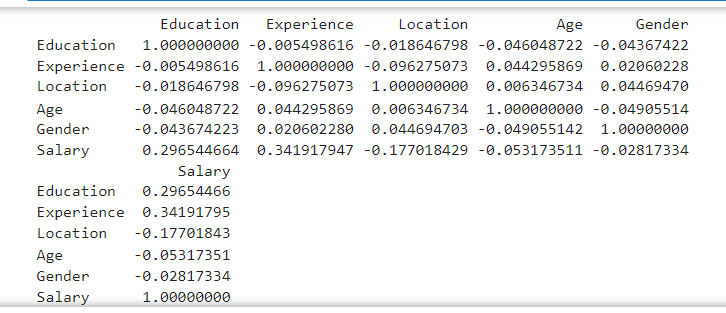
**print(correlation\_matrix)**

En aquesta anàlisi de correlació, ens centrem en com diversos factors influeixen en el salari:

Observem dues correlacions positives significatives: l'experiència amb un valor de 0,342 i l'educació amb 0,2965, la qual cosa indica que més experiència i un major nivell educatiu solen traduir-se en salaris més alts.

D'altra banda, les correlacions negatives suggereixen menys influència en el salari. La localització té la correlació més baixa (-0,177), seguida de l'edat (-0,53), que mostra que a mesura que augmenta l'edat, el salari tendeix a disminuir, i el gènere (-0,2817), que també reflecteix una influència negativa.

En resum, l'experiència i l'educació són factors positius per al salari, mentre que la localització, l'edat i el gènere tenen un impacte negatiu.

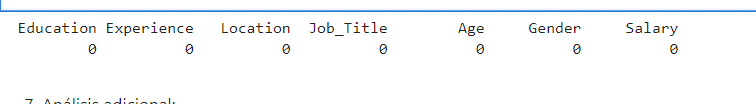


## 6. Gestió de valors perduts:

Comprovar quants valors faltants hi ha per columna

per verificar quants valors faltants (nuls) hi ha en cada columna del nostre data frame, podem utilitzar la combinació de les funcions **is.na() i colSums().**

**colSums(is.na(df))**



## 7. Anàlisi addicional:

Hem vist que l'educació és un factor important que influeix en el salari, però és fonamental saber quins factors impacten més o menys. Si hi ha algun outlier (un valor que es desvia significativament de la resta), entendre la correlació ens pot ajudar a identificar la raó d'aquell valor atípic.

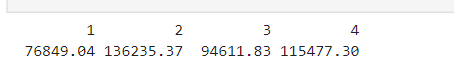
Per això, calculem la mitjana de salari per cada categoria d'educació. Això ens permet veure com varia el salari segons el nivell educatiu. Utilitzant el codi:

**# Calcular la media de salario por cada categoría en "Education"**

**mean\_salary\_by\_education <- tapply(df$Salary, df$Education, mean, na.rm = TRUE)**

# Mostrar la tabla de medias

**print(mean\_salary\_by\_education)**



Així, podrem observar les mitjanes de salari associades a cada categoria educativa.

Si trobem un outlier, és important investigar més a fons per entendre la seva causa.

Per exemple, si busquem un salari específic, com 193016.6, podem fer-ho amb el següent codi:

**tolerance <- 0.01 # Puedes ajustar la tolerancia**

**specific\_salary <- df[abs(df$Salary - 193016.6) < tolerance, ]**

**# Mostrar los resultados**

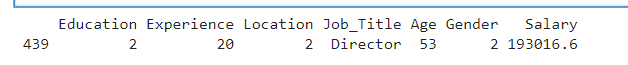
**if(nrow(specific\_salary) > 0) {**

**print(specific\_salary)**

**} else {**

**cat("No se encontró el salario 193016.6 en el DataFrame.\n")**

**}**



En el cas d’aquest outlier, té sentit que aparegui, ja que hem observat que a mesura que augmenten els anys d'experiència, també ho fa el salari. A més, pertany a la categoria de EDUCATION que té la mitjana de salari més alta.