Proyecto

Alejandro Quesada Leiva

2024-08-19

## Colegio Universitario de Cartago

## Lenguaje de Minería de Datos

## Profesor: Osvaldo Gonzalez Chaves

## II Cuatrimestre 2024

## DataSet Employees Analisys

Primero obtenemos información básica de los datos del Data Set que vamos a utilizar

class(employees)

## [1] "data.frame"

dim(employees)

## [1] 689 15

names(employees)

## [1] "No" "FirstName" "LastName" "Gender"   
## [5] "StartDate" "Years" "Department" "Country"   
## [9] "Center" "MonthlySalary" "AnnualSalary" "JobRate"   
## [13] "SickLeaves" "UnpaidLeaves" "OvertimeHours"

str(employees)

## 'data.frame': 689 obs. of 15 variables:  
## $ No : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ FirstName : chr "Ghadir" "Omar" "Ailya" "Lwiy" ...  
## $ LastName : chr "Hmshw" "Hishan" "Sharaf" "Qbany" ...  
## $ Gender : chr "Male" "Male" "Female" "Male" ...  
## $ StartDate : chr "2018-04-04" "2020-05-21" "2017-09-28" "2018-08-14" ...  
## $ Years : int 6 4 6 6 4 8 4 6 6 4 ...  
## $ Department : chr "Quality Control" "Quality Control" "Major Mfg Projects" "Manufacturing" ...  
## $ Country : chr "Egypt" "Saudi Arabia" "Saudi Arabia" "United Arab Emirates" ...  
## $ Center : chr "West" "West" "West" "Main" ...  
## $ MonthlySalary: int 1560 3247 2506 1828 970 2332 1959 3394 1479 1186 ...  
## $ AnnualSalary : int 18720 38964 30072 21936 11640 27984 23508 40728 17748 14232 ...  
## $ JobRate : num 3 1 2 3 5 3 3 5 4.5 4.5 ...  
## $ SickLeaves : int 1 0 0 0 0 3 6 0 0 1 ...  
## $ UnpaidLeaves : int 0 5 3 0 5 0 0 0 0 0 ...  
## $ OvertimeHours: int 183 198 192 7 121 8 116 7 105 153 ...

glimpse(employees)

## Rows: 689  
## Columns: 15  
## $ No <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1…  
## $ FirstName <chr> "Ghadir", "Omar", "Ailya", "Lwiy", "Ahmad", "Muhamad", "…  
## $ LastName <chr> "Hmshw", "Hishan", "Sharaf", "Qbany", "Bikri", "Zueitr",…  
## $ Gender <chr> "Male", "Male", "Female", "Male", "Male", "Male", "Femal…  
## $ StartDate <chr> "2018-04-04", "2020-05-21", "2017-09-28", "2018-08-14", …  
## $ Years <int> 6, 4, 6, 6, 4, 8, 4, 6, 6, 4, 6, 6, 5, 8, 5, 4, 8, 6, 6,…  
## $ Department <chr> "Quality Control", "Quality Control", "Major Mfg Project…  
## $ Country <chr> "Egypt", "Saudi Arabia", "Saudi Arabia", "United Arab Em…  
## $ Center <chr> "West", "West", "West", "Main", "Main", "North", "Main",…  
## $ MonthlySalary <int> 1560, 3247, 2506, 1828, 970, 2332, 1959, 3394, 1479, 118…  
## $ AnnualSalary <int> 18720, 38964, 30072, 21936, 11640, 27984, 23508, 40728, …  
## $ JobRate <dbl> 3.0, 1.0, 2.0, 3.0, 5.0, 3.0, 3.0, 5.0, 4.5, 4.5, 2.0, 1…  
## $ SickLeaves <int> 1, 0, 0, 0, 0, 3, 6, 0, 0, 1, 5, 2, 6, 1, 1, 0, 0, 4, 0,…  
## $ UnpaidLeaves <int> 0, 5, 3, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,…  
## $ OvertimeHours <int> 183, 198, 192, 7, 121, 8, 116, 7, 105, 153, 12, 70, 85, …

Realizamos la verificación de que el set no contenga nulos, en caso de que el set SI tenga nulos, se debe hacer la limpieza del set, esto con el fin de que los resultados sean los correctos.

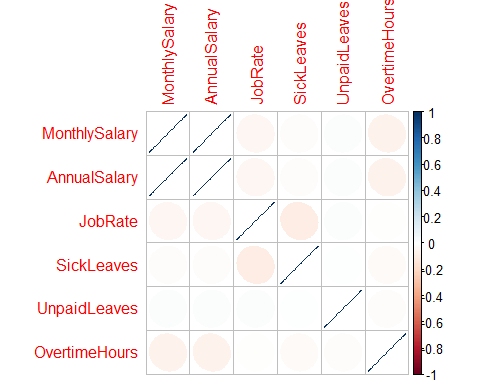
any(is.na(employees)) #false

## [1] FALSE

summary(employees)

## No FirstName LastName Gender   
## Min. : 1 Length:689 Length:689 Length:689   
## 1st Qu.:173 Class :character Class :character Class :character   
## Median :345 Mode :character Mode :character Mode :character   
## Mean :345   
## 3rd Qu.:517   
## Max. :689   
## StartDate Years Department Country   
## Length:689 Min. :3.00 Length:689 Length:689   
## Class :character 1st Qu.:4.00 Class :character Class :character   
## Mode :character Median :5.00 Mode :character Mode :character   
## Mean :5.06   
## 3rd Qu.:6.00   
## Max. :8.00   
## Center MonthlySalary AnnualSalary JobRate   
## Length:689 Min. : 703 Min. : 8436 Min. :1.000   
## Class :character 1st Qu.:1436 1st Qu.:17232 1st Qu.:3.000   
## Mode :character Median :2077 Median :24924 Median :3.000   
## Mean :2068 Mean :24818 Mean :3.586   
## 3rd Qu.:2682 3rd Qu.:32184 3rd Qu.:5.000   
## Max. :3450 Max. :41400 Max. :5.000   
## SickLeaves UnpaidLeaves OvertimeHours   
## Min. :0.00 Min. :0.0000 Min. : 0.0   
## 1st Qu.:0.00 1st Qu.:0.0000 1st Qu.: 3.0   
## Median :0.00 Median :0.0000 Median : 7.0   
## Mean :1.61 Mean :0.7591 Mean : 13.7   
## 3rd Qu.:3.00 3rd Qu.:0.0000 3rd Qu.: 10.0   
## Max. :6.00 Max. :6.0000 Max. :198.0

# Correlaciones



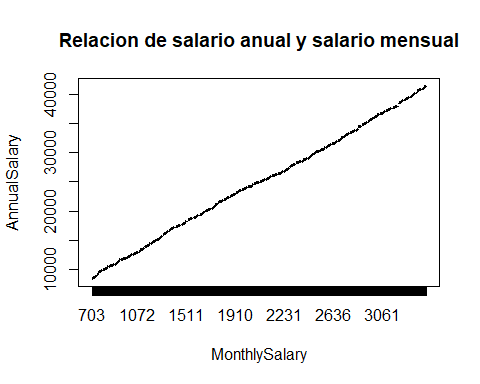
# Correlaciones encontradas (+)

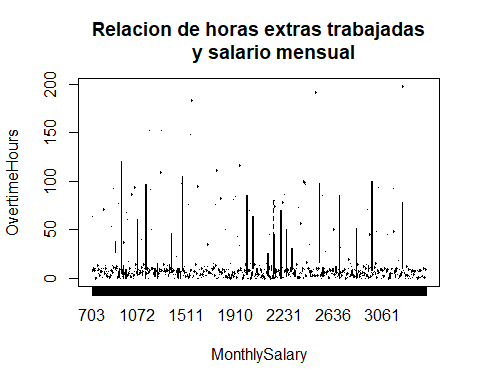
1. Annual Salary –> Monthly Salary

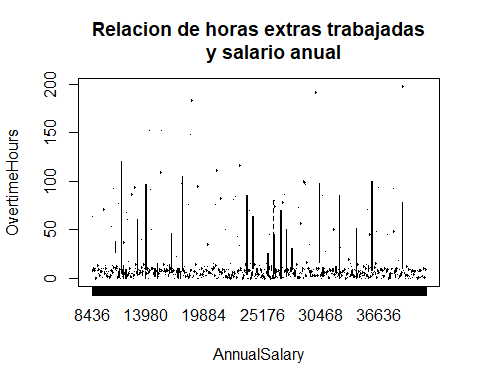
# Correlaciones encontradas (-)

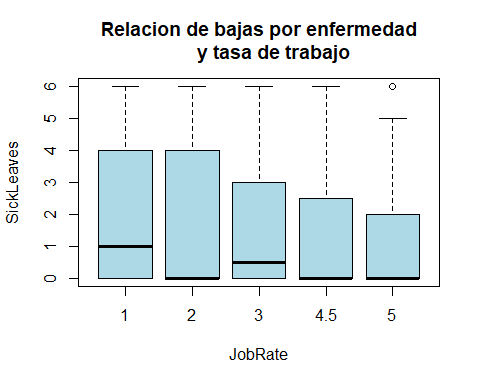
1. Overtime Hours –> Monthly Salary
2. Overtime Hours –> Annual Salary
3. Sick Leaves –> Job Rate

# Boxplot

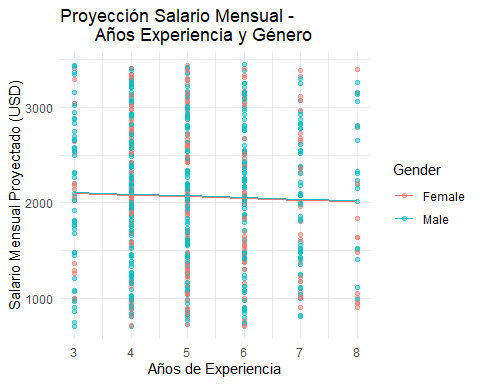








# Algoritmo (Regresion lineal multiple)



# APIs en R

# Recomendaciones:

1. Obtener datos de América o Europa, para lograr realizar una comparativa
2. Obtener el estado socioeconómico de los trabajadores (pueden existir relaciones con otros indicadores)
3. Obtener el nivel de satisfacción/felicidad de los trabajadores (pueden existir relaciones con otros indicadores)

# Conclusiones

1. El salario varía en prácticamente todos los sectores respecto al género, en unos más, en otros menos. En unos sectores ganan más las mujeres en comparación con los hombres y en otros sectores pasa lo contrario
2. Dentro de más años este laborando hay una menor probabilidad que de se tomen licencias, siendo el punto más alto a los 5 años
3. Siria y Egipto son los países con mayor probabilidad de tomar licencias por enfermedad
4. No hay relación entre salario anual y tasa de retención de los empleados, esto puede estar sujeto a otros factores como felicidad/satisfacción/relaciones etc.