Salarios en Ciencia de Datos

R-Ladies Medellín

Introduccion

En este análisis, construimos un modelo que predice los salarios de los profesionales basados en factores asociados al profesional y a la empresa.

```
library(tidyverse)  # Para la gestión y visualización de datos
library(knitr)      # Para tablas
library(broom)      # Para el resumen del modelo
library(readr)      # Para lectura de datos separados por csv
library(ggplot2)
library(hrbrthemes)
library(scales)

url <- 'https://raw.githubusercontent.com/ousuga/Datos/main/ds_salaries.csv'
salarios <- read_delim(url, delim = ";", escape_double = FALSE,
                       col_types = cols(work_year = col_character(),
                                        remote_ratio = col_character()),
                       trim_ws = TRUE)
```

Vamos a presentar los resultados del análisis de datos exploratorio en Sección y el modelo de regresión en Sección .

Análisis exploratorio de datos

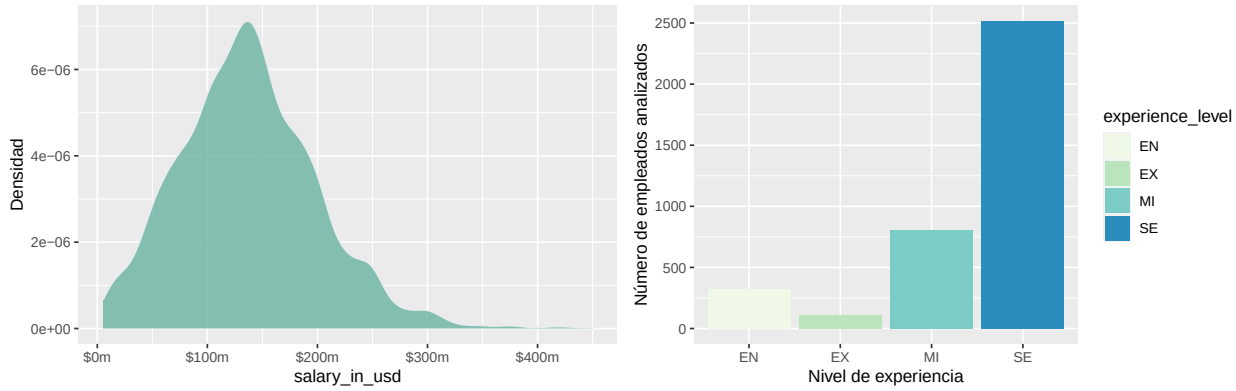
Como parte del análisis de datos exploratorios vamos a visualizar la relación entre el salario y el nivel de experiencia de los profesionales.

Visualización de datos

Figura 1 muestra la densidad de la distribución del `salary_in_usd` y un diagrama de barras del `experience_level` de los profesionales.

```
ggplot(salarios, aes(x = salary_in_usd)) +
  geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8) +
  labs(y = "Densidad") +
  scale_x_continuous(labels  = label_number(scale = 1e-3, prefix = "$",
                                            suffix = "m", accuracy = 1))

ggplot(salarios, aes(x = experience_level, fill = experience_level)) +
  geom_bar() +
  scale_fill_brewer(palette="GnBu")  +
  labs(x = "Nivel de experiencia", y = "Número de empleados analizados")
```



(a) Densidad de `salary_in_usd`

(b) Diagrama de barras de `experience_level`

Figura 1: Densidad y diagrama de barras de salario y nivel de experiencia

Figura 2 muestra la relación entre salarios y nivel de experiencia de profesionales.

```
ggplot(salarios, aes(x=salary_in_usd, group=experience_level,
                     fill=experience_level, col = experience_level)) +
  geom_density(adjust=1.5, alpha=.6)
  ylab("")
```

```
$y
[1] ""

attr(,"class")
[1] "labels"
```

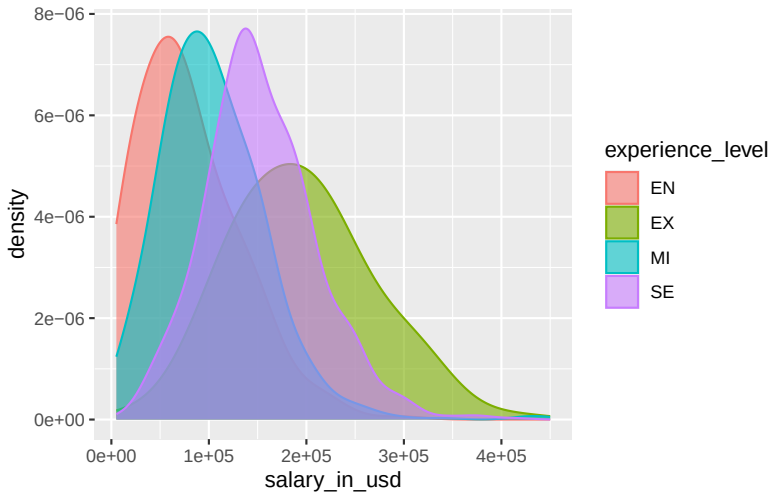


Figura 2: Salario vs Nivel de experiencia

Summary statistics

Tabla 1 muestra resumen estadístico para estas dos variables.

```
salarios %>%
  summarise(
    `Mediana de Salario` = median(salary_in_usd),
    `RIC salario` = IQR(salary_in_usd)
  ) %>%
  kable(digits = c(0, 0))
```

Tabla 1: Resumen estadístico de salarios vs Nivel de experiencia

Mediana de Salario	RIC salario
135000	80000

Modelación

Ajustamos un modelo de regresión lineal simple de la forma mostrada en la Ecuación 1.

$$Salario = \hat{\beta}_0 + \hat{\beta}_1 \times Experiencia + \epsilon \tag{1}$$

Tabla 2 muestra la salida del modelo de regresión.

```
salario_modelo <- lm(salary_in_usd ~ experience_level, data = salarios)

salario_modelo %>%
  tidy() %>%
  kable(digits = c(0, 0, 2, 2, 2))
```

Tabla 2: Modelo de regresión lineal de salarios vs nivel de experiencia

term	estimate	std.error	statistic	p.value
(Intercept)	78546	3155.80	24.89	0
experience_levelEX	116385	6157.46	18.90	0
experience_levelMI	25980	3730.68	6.96	0
experience_levelSE	74505	3350.48	22.24	0