employee.csv

2024-04-19

R Markdown

Nos conectamos al dataset:

Diccionario de Datos

Education: The educational qualifications of employees, including degree, institution, and field of study.

Joining Year: The year each employee joined the company, indicating their length of service.

City: The location or city where each employee is based or works.

Payment Tier: Categorization of employees into different salary tiers.

Age: The age of each employee, providing demographic insights.

Gender: Gender identity of employees, promoting diversity analysis.

Ever Benched: Indicates if an employee has ever been temporarily without assigned work.

Experience in Current Domain: The number of years of experience employees have in their current field.

Leave or Not: a target column

Conexión al dataset Employee.csv

El dataset "Employee.csv" se encuentra en nuestro repositorio de Github al cual nos conectamos utiliando el método read.csv.

```
url <- "https://raw.githubusercontent.com/chelocoder/ProyectoR/main/Employee.csv"
datos <- read.csv(url)
head(datos)
```

```
##   Education JoiningYear      City PaymentTier Age Gender EverBenched
## 1 Bachelors        2017 Bangalore           3  34   Male          No
## 2 Bachelors        2013      Pune           1  28 Female          No
## 3 Bachelors        2014 New Delhi           3  38 Female          No
## 4   Masters        2016 Bangalore           3  27   Male          No
## 5   Masters        2017      Pune           3  24   Male         Yes
## 6 Bachelors        2016 Bangalore           3  22   Male          No
##   ExperienceInCurrentDomain LeaveOrNot
## 1                         0          0
## 2                         3          1
## 3                         2          0
## 4                         5          1
## 5                         2          1
## 6                         0          0
```

El comando head nos permite visualizar los primeros seis registros como una vista previa que confirma la conexión y nos permite una primera vista preliminar de los datos.

Empezamos con la estadística descriptiva, esta es una rama importante de la estadística que nos permite describir las características de los datos, de ahí su nombre, pertenece a la etapa del análisis descriptivo donde empezamos a conocer los datos.

Sus partes más importantes son:

1. Estadísticos de tendencia central.
2. Estadísticos de dispersión.

1. Estadísticos Descriptivos

Obtenemos los estadísticos descriptivos de todas las variables del dataset de forma resumida:

```
# Obtener estadísticos descriptivos
summary_stats <- summary(datos)

# Mostrar estadísticos descriptivos
print(summary_stats)
```

```
##   Education          JoiningYear       City            PaymentTier
##  Length:4653        Min.   :2012   Length:4653        Min.   :1.000
##  Class :character   1st Qu.:2013   Class :character   1st Qu.:3.000
##  Mode  :character   Median :2015   Mode  :character   Median :3.000
##                     Mean   :2015                      Mean   :2.698
##                     3rd Qu.:2017                      3rd Qu.:3.000
##                     Max.   :2018                      Max.   :3.000
##       Age            Gender          EverBenched
##  Min.   :22.00   Length:4653        Length:4653
##  1st Qu.:26.00   Class :character   Class :character
##  Median :28.00   Mode  :character   Mode  :character
##  Mean   :29.39
##  3rd Qu.:32.00
##  Max.   :41.00
##  ExperienceInCurrentDomain   LeaveOrNot
##  Min.   :0.000             Min.   :0.0000
##  1st Qu.:2.000             1st Qu.:0.0000
##  Median :3.000             Median :0.0000
##  Mean   :2.906             Mean   :0.3439
##  3rd Qu.:4.000             3rd Qu.:1.0000
##  Max.   :7.000             Max.   :1.0000
```

Ahora vamos a describir cada una de las características:

1.1. Tendecia Central

Para un mejor análisis dividimos a las variables del dataset en: Variables numéricas y categóricas.

1.1.1 Variables Numéricas

Obtenemos los Estadísticos descriptivos de solamente las variables numéricas

```
##   JoiningYear    PaymentTier         Age        ExperienceInCurrentDomain
##  Min.   :2012   Min.   :1.000   Min.   :22.00   Min.   :0.000
##  1st Qu.:2013   1st Qu.:3.000   1st Qu.:26.00   1st Qu.:2.000
##  Median :2015   Median :3.000   Median :28.00   Median :3.000
##  Mean   :2015   Mean   :2.698   Mean   :29.39   Mean   :2.906
##  3rd Qu.:2017   3rd Qu.:3.000   3rd Qu.:32.00   3rd Qu.:4.000
##  Max.   :2018   Max.   :3.000   Max.   :41.00   Max.   :7.000
##    LeaveOrNot
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3439
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

Obtenemos los estadísticos descriptivos de cada variable para describir mejor las características de las variables numéricas:

a.Variable JoiningYear

a.1.Valor Mínimo:

```
# Obtener el valor mínimo
min <- min(datos$JoiningYear)

# Imprimir el valor mínimo de la columna
print(min)
```

```
## [1] 2012
```

El menor año en que se contrató a un empleado fue el año 2012.

a.2.Primer Cuartil:

```
# Obtener el primer cuartil
q1 <- quantile(datos$JoiningYear, 0.25)

# Imprimir el primer cuartil
print(q1)
```

```
##  25%
## 2013
```

El año 2013 es igual o mayor al 25% de los años e igual o menor al 75% de los años de contratación.

a.3. Mediana:

```
# Calcular la mediana de la variable
mediana_joining_year <- median(datos$JoiningYear)

# Imprimir la mediana de la variable
print(mediana_joining_year)
```

```
## [1] 2015
```

El año 2015 es igual o mayor al 50% de los años e igual o menor al 50% de los años de contratación.

a.4.Tercer Cuartil:

```
# Obtener el tercer cuartil
q3 <- quantile(datos$JoiningYear, 0.75)

# Imprimir el tercer cuartil
print(q3)
```

```
##  75%
## 2017
```

El año 2017 es igual o mayor al 75% de los años e igual o menor al 25% de los años de contratación.

a.5.Valor Máximo:

```
# Calcular el valor máximo de la variable
max <- max(datos$JoiningYear)

# Imprimir el valor máximo de la variable
print(max)
```

```
## [1] 2018
```

El mayor año en que se contrató a un empleado fue el año 2018.

a.6.Recorrido (Rango):

```
# Calcular el rango de la variable
rango <- max(datos$JoiningYear) - min(datos$JoiningYear)

# Imprimir el rango de la variable "JoiningYear"
print(rango)
```

```
## [1] 6
```

La diferencia entre el valor mayor y el valor menor es 6 (rango o recorrido).

a.7.Moda:

```
# Calcular la moda de la variable
moda <- as.numeric(names(sort(-table(datos$JoiningYear))[1]))

# Imprimir la moda de la variable
print(moda)
```

```
## [1] 2017
```

La moda es 2017 que es el valor que más se repite, es decir el que presenta la frecuencia más alta.

a.8.Frecuencia Mayor:

```r
# Calcular la frecuencia de la variable
frecuencia <- table(datos$JoiningYear)

# Imprimir la frecuencia de la variable
print(frecuencia)
```

```
##
## 2012 2013 2014 2015 2016 2017 2018
##  504  669  699  781  525 1108  367
```

```r
#Imprimir la frecuencia máxima
frecuencia_maxima <- max(frecuencia)
print(frecuencia_maxima)
```

```
## [1] 1108
```

La frecuencia mayor es 1108 que determina la moda, es decir el valor que más se repite.

a.9.Media:

```r
media <- mean(datos$JoiningYear)
print(media)
```

```
## [1] 2015.063
```

La media es 2015.063, es decir si la distirbución es simétrica y no hay valores atípicos, en promedio, los empleados de la empresa son contratados el año 2015, sin embargo, si la distirbución es asimétrica o hay valores atípicos una medida más representativa sería la mediana.

b.Variable PaymentTier

b.1.Valor Mínimo:

```r
# Obtener el valor mínimo
min <- min(datos$PaymentTier)

# Imprimir el valor mínimo de la columna
print(min)
```

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

El menor código de categoría de pagos es 1.

b.2.Primer Cuartil:

```r
# Obtener el primer cuartil
q1 <- quantile(datos$PaymentTier, 0.25)

# Imprimir el primer cuartil
print(q1)
```

```
## 25%
##   3
```

El valor 3 es igual o mayor al 25% de las categorías de pago e igual o menor al 75% de las categorías de pago.

b.3. Mediana:

```
# Calcular la mediana de la variable
mediana <- median(datos$PaymentTier)

# Imprimir la mediana de la variable
print(mediana)
```

```
## [1] 3
```

El valor 3 es igual o mayor al 50% de las categorías de pago e igual o menor al 50% de las categorías de pago.

b.4.Tercer Cuartil:

```
# Obtener el tercer cuartil
q3 <- quantile(datos$PaymentTier, 0.75)

# Imprimir el tercer cuartil
print(q3)
```

```
## 75%
##   3
```

El valor 3 es igual o mayor al 75% de las categorías de pago e igual o menor al 25% de las categorías de pago.

b.5.Valor Máximo:

```
# Calcular el valor máximo de la variable
max <- max(datos$PaymentTier)

# Imprimir el valor máximo de la variable
print(max)
```

```
## [1] 3
```

El mayor código de categoría de pagos es 3.

b.6.Recorrido (Rango):

```
# Calcular el rango de la variable
rango <- max(datos$PaymentTier) - min(datos$PaymentTier)

# Imprimir el rango de la variable
print(rango)
```

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

La diferencia entre el valor mayor y el valor menor es 2 (rango o recorrido).

b.7.Moda:

```r
# Calcular la moda de la variable
moda <- as.numeric(names(sort(-table(datos$PaymentTier))[1]))

# Imprimir la moda de la variable
print(moda)
```

```
## [1] 3
```

La moda es 3 que es el valor que más se repite, es decir el que presenta la frecuencia más alta.

b.8.Frecuencia:

```r
# Calcular la frecuencia de la variable
frecuencia <- table(datos$PaymentTier)

# Imprimir la frecuencia de la variable
print(frecuencia)
```

```
##
##    1    2    3
##  243  918 3492
```

```r
#Imprimir la frecuencia máxima
frecuencia_maxima <- max(frecuencia)
print(frecuencia_maxima)
```

```
## [1] 3492
```

La frecuencia mayor es 3492 que determina la moda, es decir el valor que más se repite.

b.9.Media:

```r
media <- mean(datos$PaymentTier)
print(media)
```

```
## [1] 2.698259
```

La media es 2.698259, es decir si la distribución es simétrica y no hay valores atípicos, en promedio, los empleados de la empresa se agrupan cerca de la categoría 3, sin embargo, si la distribución es asimétrica o hay valores atípicos una medida más representativa sería la mediana.

c.Variable Age

c.1.Valor Mínimo:

```r
# Obtener el valor mínimo
min <- min(datos$Age        )

# Imprimir el valor mínimo de la columna
print(min)
```

```
## [1] 22
```

La menor edad de un empleado es 22 años.

c.2.Primer Cuartil:

```
# Obtener el primer cuartil
q1 <- quantile(datos$Age        , 0.25)

# Imprimir el primer cuartil
print(q1)
```

```
## 25%
##  26
```

La edad 26 es igual o mayo al 25% de las edades de los empleados e igual o menor al 75% de las edades de los empleados.

c.3. Mediana:

```
# Calcular la mediana de la variable
mediana <- median(datos$Age)

# Imprimir la mediana de la variable
print(mediana)
```

```
## [1] 28
```

La edad 28 es igual o mayo al 50% de las edades de los empleados e igual o menor al 50% de las edades de los empleados.

c.4.Tercer Cuartil:

```
# Obtener el tercer cuartil
q3 <- quantile(datos$Age, 0.75)

# Imprimir el tercer cuartil
print(q3)
```

```
## 75%
##  32
```

La edad 32 es igual o mayo al 75% de las edades de los empleados e igual o menor al 25% de las edades de los empleados.

c.5.Valor Máximo:

```
# Calcular el valor máximo de la variable
max <- max(datos$Age        )

# Imprimir el valor máximo de la variable
print(max)
```

```
## [1] 41
```

La mayor edad de un empleado es 41 años.

c.6.Recorrido (Rango):

```
# Calcular el rango de la variable
rango <- max(datos$Age) - min(datos$Age)

# Imprimir el rango de la variable
print(rango)
```

```
## [1] 19
```

La diferencia entre el valor mayor de las edades y el valor menor es 19 (rango o recorrido).

c.7.Moda:

```
# Calcular la moda de la variable
moda <- as.numeric(names(sort(-table(datos$Age))[1]))

# Imprimir la moda de la variable
print(moda)
```

```
## [1] 26
```

La moda es 26 que es el valor que más se repite, es decir el que presenta la frecuencia más alta.

c.8.Frecuencia:

```
# Calcular la frecuencia de la variable
frecuencia <- table(datos$Age)

# Imprimir la frecuencia de la variable
print(frecuencia)
```

```
##
##  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41
##  49  48 385 418 645 625 630 230 220 125 132 124 136 123 139 141 136 131 134  82
```

```
#Imprimir la frecuencia máxima
frecuencia_maxima <- max(frecuencia)
print(frecuencia_maxima)
```

```
## [1] 645
```

La frecuencia mayor es 645 que determina la moda, es decir el valor que más se repite.

c.9.Media:

```
media <- mean(datos$Age)
print(media)
```

```
## [1] 29.39329
```

La media es 29.39329, es decir si la distribución es simétrica y no hay valores atípicos, en promedio, los empleados de la empresa se agrupan alrededor de la edad de 29 años, sin embargo, si la distribución es asimétrica o hay valores atípicos una medida más representativa sería la mediana.

d.Variable ExperienceInCurrentDomain

d.1.Valor Mínimo:

```
# Obtener el valor mínimo
min <- min(datos$ExperienceInCurrentDomain)

# Imprimir el valor mínimo de la columna
print(min)
```

```
## [1] 0
```

El menor valor en años de experiencia de un empleado es 0.

d.2.Primer Cuartil:

```
# Obtener el primer cuartil
q1 <- quantile(datos$ExperienceInCurrentDomain, 0.25)

# Imprimir el primer cuartil
print(q1)
```

```
## 25%
##   2
```

El valor 2 años de experiencia de un empleado es igual o mayor al 25% de los años de experiencia e igual o menor al 75% de los años de experiencia de otros empleados de la empresa.

d.3. Mediana:

```
# Calcular la mediana de la variable
mediana <- median(datos$ExperienceInCurrentDomain)

# Imprimir la mediana de la variable
print(mediana)
```

```
## [1] 3
```

El valor 3 años de experiencia de un empleado es igual o mayor al 50% de los años de experiencia e igual o menor al 50% de los años de experiencia de otros empleados de la empresa.

d.4.Tercer Cuartil:

```
# Obtener el tercer cuartil
q3 <- quantile(datos$ExperienceInCurrentDomain, 0.75)

# Imprimir el tercer cuartil
print(q3)
```

```
## 75%
##   4
```

El valor 4 años de experiencia de un empleado es igual o mayor al 75% de los años de experiencia e igual o menor al 25% de los años de experiencia de otros empleados de la empresa.

d.5.Valor Máximo:

```
# Calcular el valor máximo de la variable
max <- max(datos$ExperienceInCurrentDomain)

# Imprimir el valor máximo de la variable
print(max)
```

```
## [1] 7
```

El mayor valor en años de experiencia de un empleado es 7.

d.6.Recorrido (Rango):

```
# Calcular el rango de la variable
rango <- max(datos$ExperienceInCurrentDomain) - min(datos$ExperienceInCurrentDomain)

# Imprimir el rango de la variable
print(rango)
```

```
## [1] 7
```

La diferencia entre el mayor valor en años de experiencia y el valor menor es 7 (rango o recorrido).

d.7.Moda:

```
# Calcular la moda de la variable
moda <- as.numeric(names(sort(-table(datos$ExperienceInCurrentDomain))[1]))

# Imprimir la moda de la variable
print(moda)
```

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

La moda es 2 años de experiencia que es el valor que más se repite, es decir el que presenta la frecuencia más alta.

d.8.Frecuencia:

```
# Calcular la frecuencia de la variable
frecuencia <- table(datos$ExperienceInCurrentDomain)

# Imprimir la frecuencia de la variable
print(frecuencia)
```

```
##
##    0    1    2    3    4    5    6    7
##  355  558 1087  786  931  919    8    9
```

```r
#Imprimir la frecuencia máxima
frecuencia_maxima <- max(frecuencia)
print(frecuencia_maxima)
```

```
## [1] 1087
```

La frecuencia mayor es 1087 que determina la moda, es decir el valor que más se repite.

d.9.Media:

```r
media <- mean(datos$ExperienceInCurrentDomain)
print(media)
```

```
## [1] 2.905652
```

La media es 2.905652, es decir si la distribución es simétrica y no hay valores atípicos, en promedio, los empleados de la empresa tienen alrededor de 3 años de experiencia, sin embargo, si la distribución es asimétrica o hay valores atípicos una medida más representativa sería la mediana.

e.Variable LeaveOrNot (Variable objetivo)

e.1.Valor Mínimo:

```r
# Obtener el valor mínimo
min <- min(datos$LeaveOrNot)

# Imprimir el valor mínimo de la columna
print(min)
```

```
## [1] 0
```

El menor código es 0 que representa a "No renunció".

e.2.Primer Cuartil:

```r
# Obtener el primer cuartil
q1 <- quantile(datos$LeaveOrNot, 0.25)

# Imprimir el primer cuartil
print(q1)
```

```
## 25%
##   0
```

Los que no renunciaron son igual o mayor al 25% de los valores objetivo e igual o menor al 75% de los valores objetivo.

e.3. Mediana:

```r
# Calcular la mediana de la variable
mediana <- median(datos$LeaveOrNot)

# Imprimir la mediana de la variable
print(mediana)
```

```
## [1] 0
```

Los que no renunciaron son igual o mayor al 50% de los valores objetivo e igual o menor al 50% de los valores objetivo.

e.4.Tercer Cuartil:

```r
# Obtener el tercer cuartil
q3 <- quantile(datos$LeaveOrNot, 0.75)

# Imprimir el tercer cuartil
print(q3)
```

```
## 75%
##   1
```

Los que sí renunciaron son igual o mayor al 75% de los valores objetivo e igual o menor al 25% de los valores objetivo.

e.5.Valor Máximo:

```r
# Calcular el valor máximo de la variable
max <- max(datos$LeaveOrNot)

# Imprimir el valor máximo de la variable
print(max)
```

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

El menor código es 1 que representa a "Sí renunció".

e.6.Recorrido (Rango):

```r
# Calcular el rango de la variable
rango <- max(datos$LeaveOrNot) - min(datos$LeaveOrNot)

# Imprimir el rango de la variable
print(rango)
```

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

La diferencia entre el valor mayor y el valor menor es 1 (rango o recorrido).

e.7.Moda:

```r
# Calcular la moda de la variable
moda <- as.numeric(names(sort(-table(datos$LeaveOrNot))[1]))

# Imprimir la moda de la variable
print(moda)
```

```
## [1] 0
```

La moda es 0 (Los que no renunciaron) que es el valor que más se repite, es decir el que presenta la frecuencia más alta.

e.8.Frecuencia:

```
# Calcular la frecuencia de la variable
frecuencia <- table(datos$LeaveOrNot)

# Imprimir la frecuencia de la variable
print(frecuencia)
```

```
##
##    0    1
## 3053 1600
```

```
#Imprimir la frecuencia máxima
frecuencia_maxima <- max(frecuencia)
print(frecuencia_maxima)
```

```
## [1] 3053
```

La frecuencia mayor es 3053 que determina la moda, es decir el valor que más se repite.

e.9.Media:

```
media <- mean(datos$LeaveOrNot)
print(media)
```

```
## [1] 0.3438642
```

La media es 0.3438642, es decir si la distribución es simétrica y no hay valores atípicos, en promedio, los empleados de la empresa se agrupan alreder del valor 0 (Sí renunciaron), sin embargo, si la distribución es asimétrica o hay valores atípicos una medida más representativa sería la mediana.

1.1.2 Variables Categóricas

Obtenemos los estadísticos descriptivos de las variables categóri, aquellas que no están representadas de forma numérica:

```
# Seleccionar solo las columnas categóricas
categorical_data <- datos %>%
  select_if(is.character)

# Obtener estadísticos descriptivos para las variables categóricas
summary_stats_categorical <- summary(categorical_data)

# Imprimir los estadísticos descriptivos
print(summary_stats_categorical)
```

```
##   Education             City              Gender          EverBenched
##  Length:4653        Length:4653        Length:4653        Length:4653
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

Obtenemos los estadísticos descriptivos de cada variable:

a.Variable Education

a.1.Frecuencia:

```
# Calcular la frecuencia de cada valor en la variable categórica
frecuencia <- table(datos$Education)

# Imprimir los valores y sus frecuencias
print(frecuencia)
```

```
##
## Bachelors   Masters       PHD
##      3601       873       179
```

La frecuencia mayor es 3601 que determina la moda, es decir el valor que más se repite.

a.2.Moda:

```
# Calcular la frecuencia de cada valor en la variable categórica
frecuencia <- table(datos$Education)

# Encontrar el valor que más se repite
valor_mas_repetido <- names(frecuencia)[which.max(frecuencia)]

# Imprimir el valor que más se repite
print(valor_mas_repetido)
```

```
## [1] "Bachelors"
```

La moda es “Bachelors” que es el valor que más se repite, es decir el que presenta la frecuencia más alta.

b.Variable City

b.1.Frecuencia:

```
# Calcular la frecuencia de cada valor en la variable categórica
frecuencia <- table(datos$City)

# Imprimir los valores y sus frecuencias
print(frecuencia)
```

```
##
## Bangalore New Delhi      Pune
##      2228      1157      1268
```

La frecuencia mayor es 2228 que determina la moda, es decir el valor que más se repite.

b.2.Moda:

```r
# Calcular la frecuencia de cada valor en la variable categórica
frecuencia <- table(datos$City)

# Encontrar el valor que más se repite
valor_mas_repetido <- names(frecuencia)[which.max(frecuencia)]

# Imprimir el valor que más se repite
print(valor_mas_repetido)
```

```
## [1] "Bangalore"
```

La moda es “Bangalore” que es el valor que más se repite, es decir el que presenta la frecuencia más alta.

c.Variable Gender

c.1.Frecuencia:

```r
# Calcular la frecuencia de cada valor en la variable categórica
frecuencia <- table(datos$Gender)

# Imprimir los valores y sus frecuencias
print(frecuencia)
```

```
##
## Female   Male
##   1875   2778
```

La frecuencia mayor es 2778 que determina la moda, es decir el valor que más se repite.

c.2.Moda:

```r
# Calcular la frecuencia de cada valor en la variable categórica
frecuencia <- table(datos$Gender)

# Encontrar el valor que más se repite
valor_mas_repetido <- names(frecuencia)[which.max(frecuencia)]

# Imprimir el valor que más se repite
print(valor_mas_repetido)
```

```
## [1] "Male"
```

La moda es “Male” que es el valor que más se repite, es decir el que presenta la frecuencia más alta.

d.Variable EverBenched

d.1.Frecuencia:

```r
# Calcular la frecuencia de cada valor en la variable categórica
frecuencia <- table(datos$EverBenched)

# Imprimir los valores y sus frecuencias
print(frecuencia)
```

```
##
##   No  Yes
## 4175  478
```

La frecuencia mayor es 4175 que determina la moda, es decir el valor que más se repite.

d.2.Moda:

```r
# Calcular la frecuencia de cada valor en la variable categórica
frecuencia <- table(datos$EverBenched)

# Encontrar el valor que más se repite
valor_mas_repetido <- names(frecuencia)[which.max(frecuencia)]

# Imprimir el valor que más se repite
print(valor_mas_repetido)
```

```
## [1] "No"
```

La moda es "No" que es el valor que más se repite, es decir el que presenta la frecuencia más alta.

1.2. Dispersión

Las medidas de dispersión nos indican el grado de variabilidad de los valores, si el grado de variabilidad o dispersión es significativo entonces una mejor medida sería la mediana, en cambio cuando la dispersión o variabilidad de los valores no es no es significativa una mejor medida sería la media.

1.1.1 Variables Numéricas

```r
# Seleccionar solo las columnas numéricas
numeric_data <- datos %>%
  select_if(is.numeric)

# Calcular el rango
rango <- apply(numeric_data, 2, function(x) max(x) - min(x))

# Calcular la varianza
varianza <- apply(numeric_data, 2, var)

# Calcular la desviación estándar
desviacion_estandar <- apply(numeric_data, 2, sd)

# Crear un dataframe con los resultados
estadisticos_dispersion <- data.frame(
  Variable = names(numeric_data),
  Rango = rango,
  Varianza = varianza,
  Desviacion_Estandar = desviacion_estandar
)

# Imprimir los estadísticos de dispersión
print(estadisticos_dispersion)
```

```
##                                         Variable Rango   Varianza
## JoiningYear                           JoiningYear     6  3.4721732
## PaymentTier                           PaymentTier     2  0.3152098
## Age                                           Age    19 23.2911158
## ExperienceInCurrentDomain ExperienceInCurrentDomain     7  2.4281129
## LeaveOrNot                             LeaveOrNot     1  0.2256701
##                           Desviacion_Estandar
## JoiningYear                         1.8633768
## PaymentTier                         0.5614355
## Age                                 4.8260870
## ExperienceInCurrentDomain           1.5582403
## LeaveOrNot                          0.4750475
```

Se puede apreciar que las variables Age tiene la mayor varianza lo que puede indicar la presencia de valores atípicos (outliers).

2. Análisis Exploratorio de Datos (EDA)

```r
# Conteo de datos faltantes en cada columna
missing_data_counts <- sapply(datos, function(x) sum(is.na(x)))

# Mostrando el conteo de datos faltantes
missing_data_counts
```

```
##                 Education                JoiningYear                       City
##                         0                          0                          0
##               PaymentTier                        Age                     Gender
##                         0                          0                          0
##               EverBenched ExperienceInCurrentDomain                 LeaveOrNot
##                         0                          0                          0
```

Como contamos con pocas variables, se puede destacar a primera vista que tenemos 8 variables independientes y 1 dependiente. De las variables independientes, 4 son categóricas y 4 son numéricas. La variable dependiente es numérica.

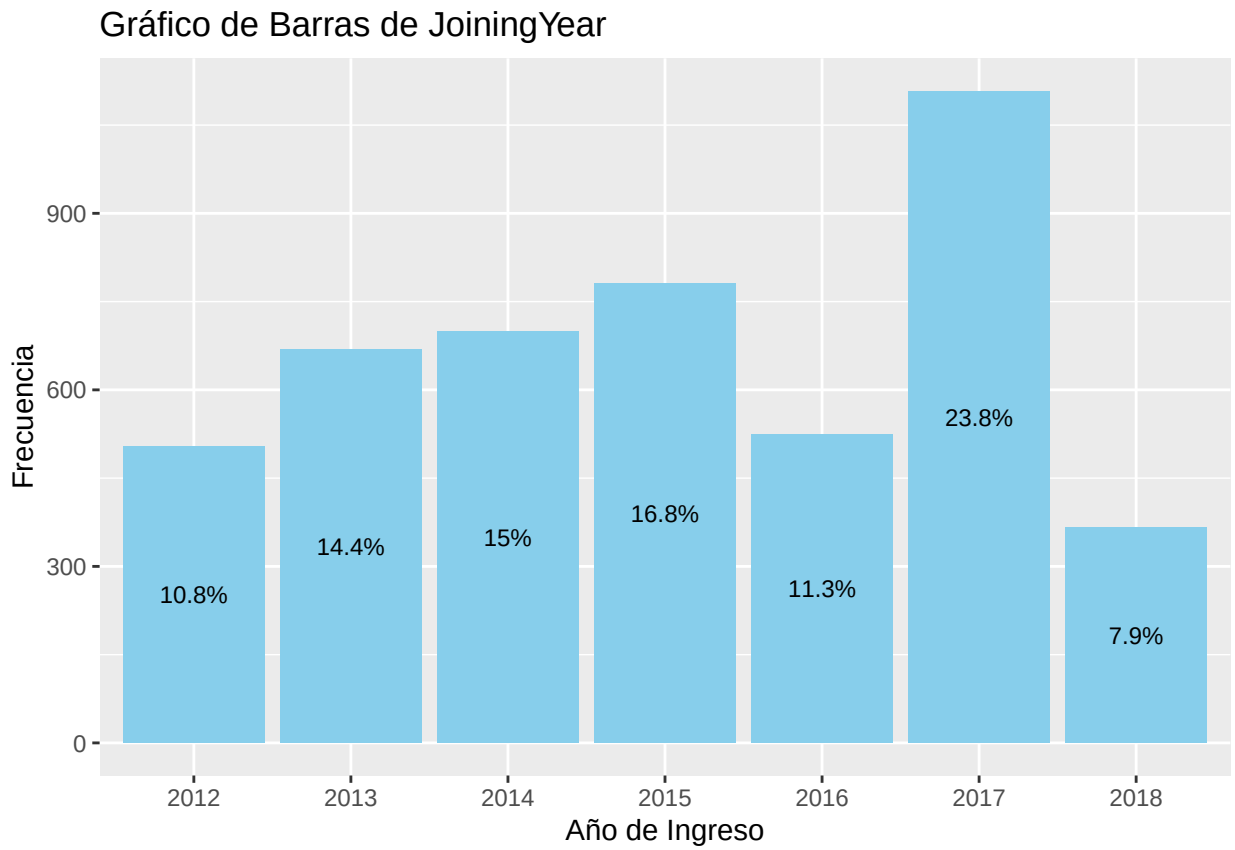
```r
# Identificando variables numéricas
numeric_vars <- sapply(datos, function(x) is.numeric(x) || is.integer(x))
numeric_variables <- names(datos)[numeric_vars]

# Identificando variables categóricas
categorical_vars <- sapply(datos, function(x) is.factor(x) || is.character(x) || is.logical(x))
categorical_variables <- names(datos)[categorical_vars]

# Mostrando las variables numéricas y categóricas
list(numeric = numeric_variables, categorical = categorical_variables)
```

```
## $numeric
## [1] "JoiningYear"               "PaymentTier"
## [3] "Age"                       "ExperienceInCurrentDomain"
## [5] "LeaveOrNot"
##
```

```
## $categorical
## [1] "Education"   "City"        "Gender"      "EverBenched"
```

Decidimos realizar un gráfico de barras de JoiningYear, debido a que su rango de valores era corto. Podemos observar que el año que tuvo más personas ingresando a trabajar fue el 2017 (23.8%) y el año con menos ingresos fue el 2018 (7.9%).

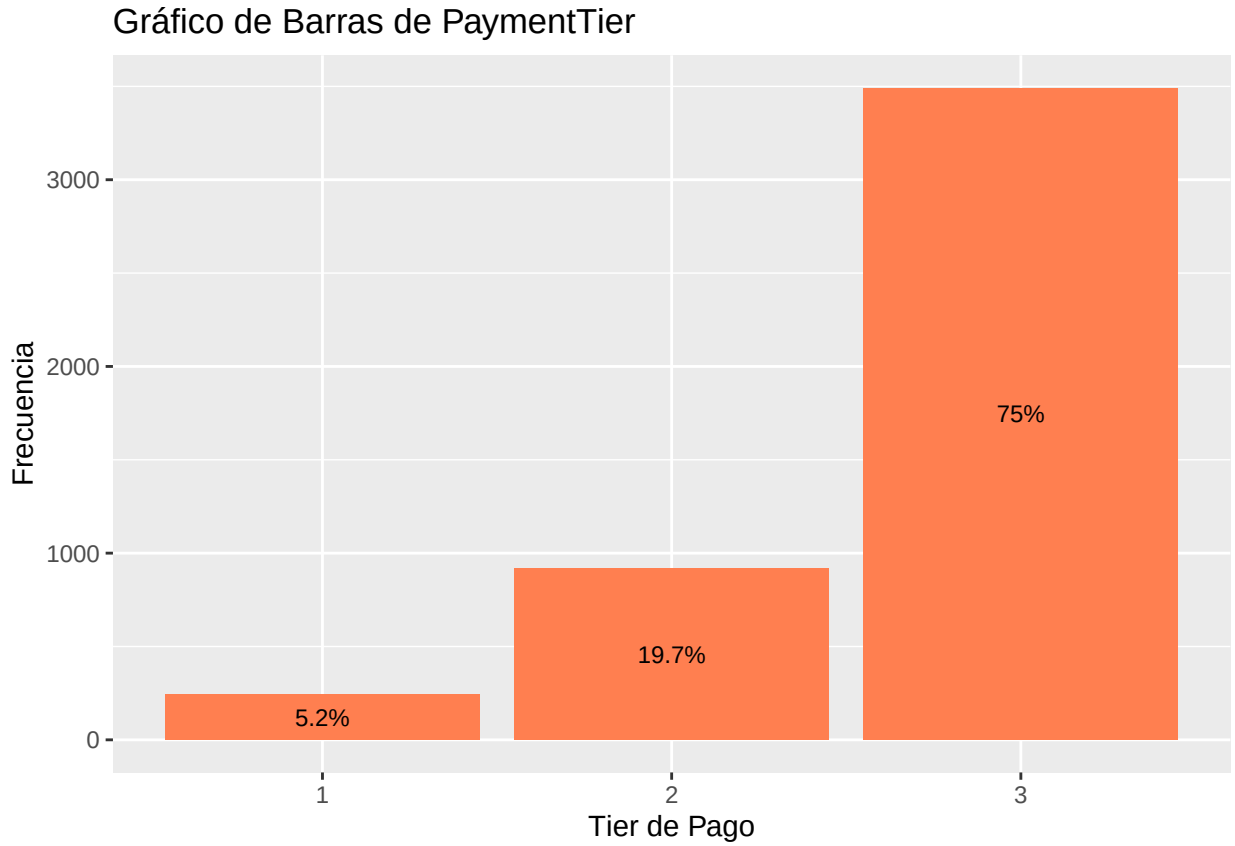
```
# Calcular los porcentajes de frecuencia de cada año
library(ggplot2)
joining_year_data <- as.data.frame(table(datos$JoiningYear))
joining_year_data$Percentage <- joining_year_data$Freq / sum(joining_year_data$Freq) * 100

# Crear el gráfico de barras con porcentajes usando ggplot2
ggplot(joining_year_data, aes(x = Var1, y = Freq, label = paste0(round(Percentage, 1), "%"))) +
  geom_bar(stat = "identity", fill = "skyblue") +
  geom_text(size = 3, position = position_stack(vjust = 0.5), color = "black") +
  ggtitle("Gráfico de Barras de JoiningYear") +
  xlab("Año de Ingreso") +
  ylab("Frecuencia")
```



Decidimos realizar un gráfico de barras de PaymentTier, debido a que su rango de valores era corto. Podemos observar que el Tier 3, que precisamente corresponde al salario más bajo, es el más frecuente (75%), mientras que el Tier 1, donde se gana más, es el menos frecuente (5.2%).

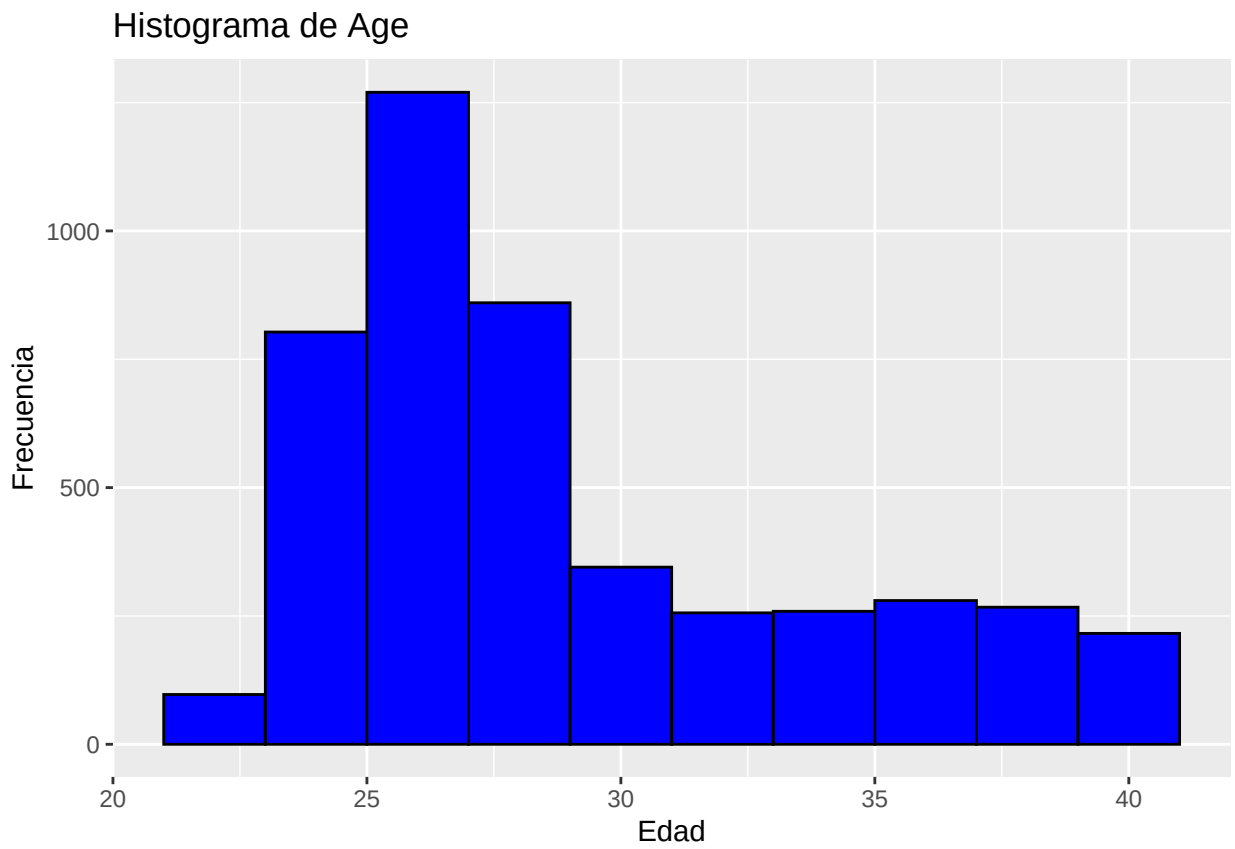
```
# Calcular los porcentajes de frecuencia de cada categoría de PaymentTier
library(ggplot2)
```

```r
payment_tier_data <- as.data.frame(table(datos$PaymentTier))
payment_tier_data$Percentage <- payment_tier_data$Freq / sum(payment_tier_data$Freq) * 100

# Crear el gráfico de barras con porcentajes usando ggplot2
ggplot(payment_tier_data, aes(x = Var1, y = Freq, label = paste0(round(Percentage, 1), "%"))) +
  geom_bar(stat = "identity", fill = "coral") +
  geom_text(size = 3, position = position_stack(vjust = 0.5), color = "black") +
  ggtitle("Gráfico de Barras de PaymentTier") +
  xlab("Tier de Pago") +
  ylab("Frecuencia")
```



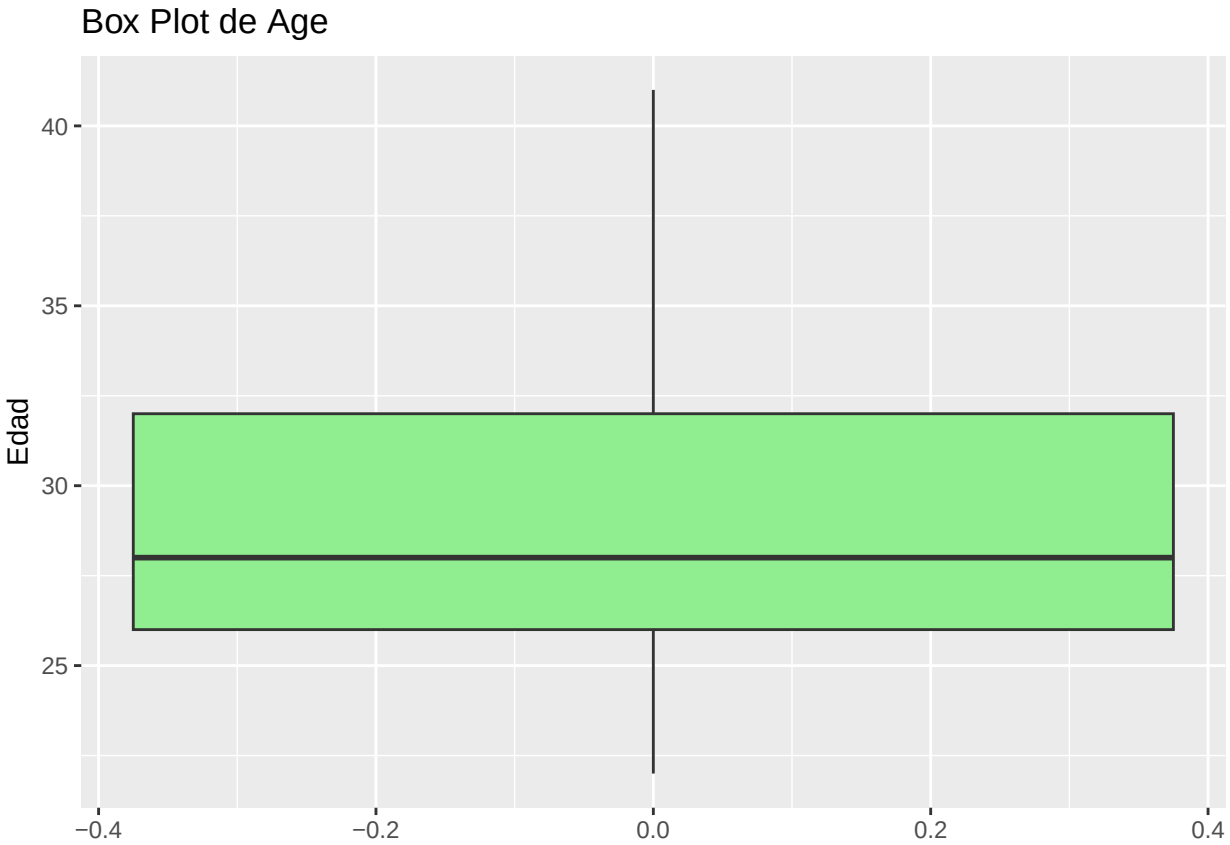
Decidimos realizar un histograma de Age, debido a que su rango de valores era más amplio. Podemos observar que hay mayor cantidad de personas que tienen entre 25 y 27 años, y la menor cantidad de personas son las que tienen entre 21 y 23 años. Ademas la data esta sesgada a la derecha.

```r
# Cargar la librería ggplot2
library(ggplot2)

# Crear un histograma usando ggplot2
ggplot(datos, aes(x = Age)) +
  geom_histogram(binwidth = 2, fill="blue", color="black") +
  ggtitle("Histograma de Age") +
  xlab("Edad") +
  ylab("Frecuencia")
```



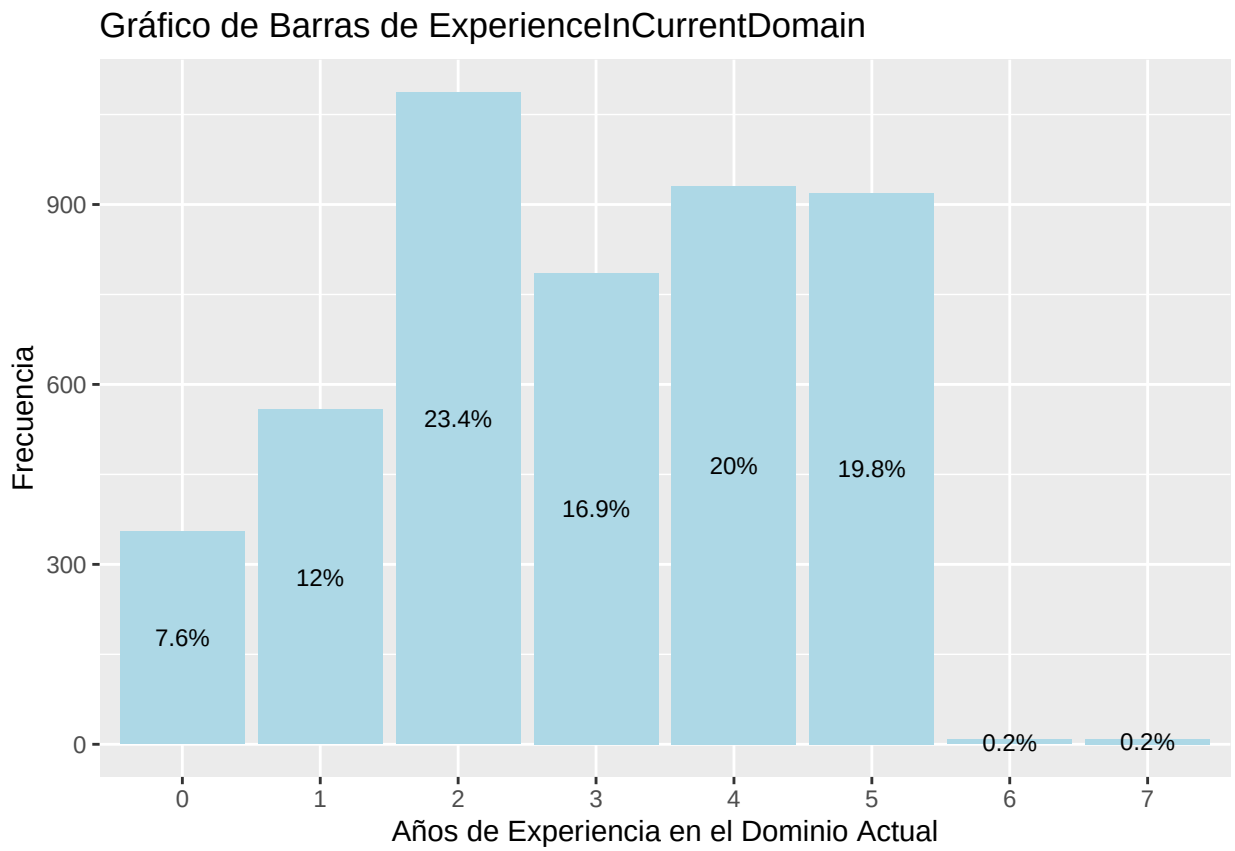
Adicionalmente, al analizar Age con un boxplot, podemos ver que no hay presencia de outliers.

```
# Cargar la librería ggplot2
library(ggplot2)

# Creando un box plot usando ggplot2
ggplot(datos, aes(y = Age)) +
  geom_boxplot(fill = "lightgreen") +
  ggtitle("Box Plot de Age") +
  ylab("Edad")
```



Decidimos realizar un gráfico de barras de ExperienceInCurrentDomain, debido a que su rango de valores era corto. Podemos observar que la frecuencia más alta es cuando las personas tienen 2 años de experiencia (23.4%), mientras que la más baja es cuando tienen 6 o 7 años de experiencia (0.2% cada uno).

```
# Calcular los porcentajes de frecuencia de cada valor de ExperienceInCurrentDomain
library(ggplot2)
experience_data <- as.data.frame(table(datos$ExperienceInCurrentDomain))
experience_data$Percentage <- experience_data$Freq / sum(experience_data$Freq) * 100

# Crear el gráfico de barras con porcentajes usando ggplot2
ggplot(experience_data, aes(x = Var1, y = Freq, label = paste0(round(Percentage, 1), "%"))) +
  geom_bar(stat = "identity", fill = "lightblue") +
  geom_text(size = 3, position = position_stack(vjust = 0.5), color = "black") +
  ggtitle("Gráfico de Barras de ExperienceInCurrentDomain") +
  xlab("Años de Experiencia en el Dominio Actual") +
  ylab("Frecuencia")
```



Ahora, analizando nuestra variable dependiente, decidimos realizar un gráfico de barras de LeaveOrNot, debido a que su rango de valores era corto. Podemos observar que la frecuencia más alta es que las personas se queden (65.6%); sin embargo, las personas que se van también representan una parte considerable (34.4%).

```
# Calcular los porcentajes de frecuencia de cada categoría de LeaveOrNot
library(ggplot2)
leave_data <- as.data.frame(table(datos$LeaveOrNot))
leave_data$Percentage <- leave_data$Freq / sum(leave_data$Freq) * 100

# Crear el gráfico de barras con porcentajes usando ggplot2
ggplot(leave_data, aes(x = Var1, y = Freq, label = paste0(round(Percentage, 1), "%"))) +
  geom_bar(stat = "identity", fill = "salmon") +
  geom_text(size = 3, position = position_stack(vjust = 0.5), color = "black") +
  ggtitle("Gráfico de Barras de LeaveOrNot") +
  xlab("Decisión") +
  ylab("Frecuencia")
```

```
# Identifying character variables
char_vars <- sapply(datos, class) == "character"

# Extracting unique values for each character variable
unique_values <- lapply(datos[, char_vars, drop = FALSE], unique)

# Displaying the unique values
unique_values
```

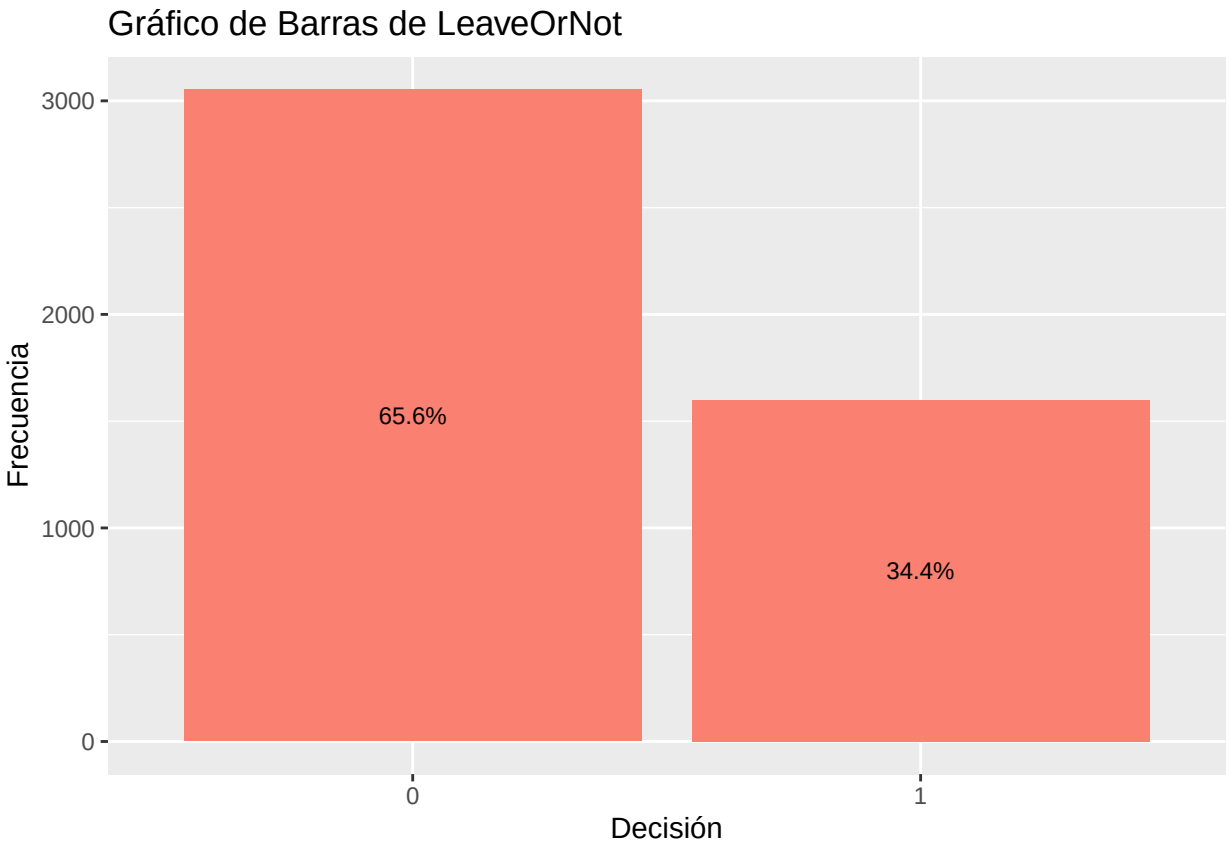
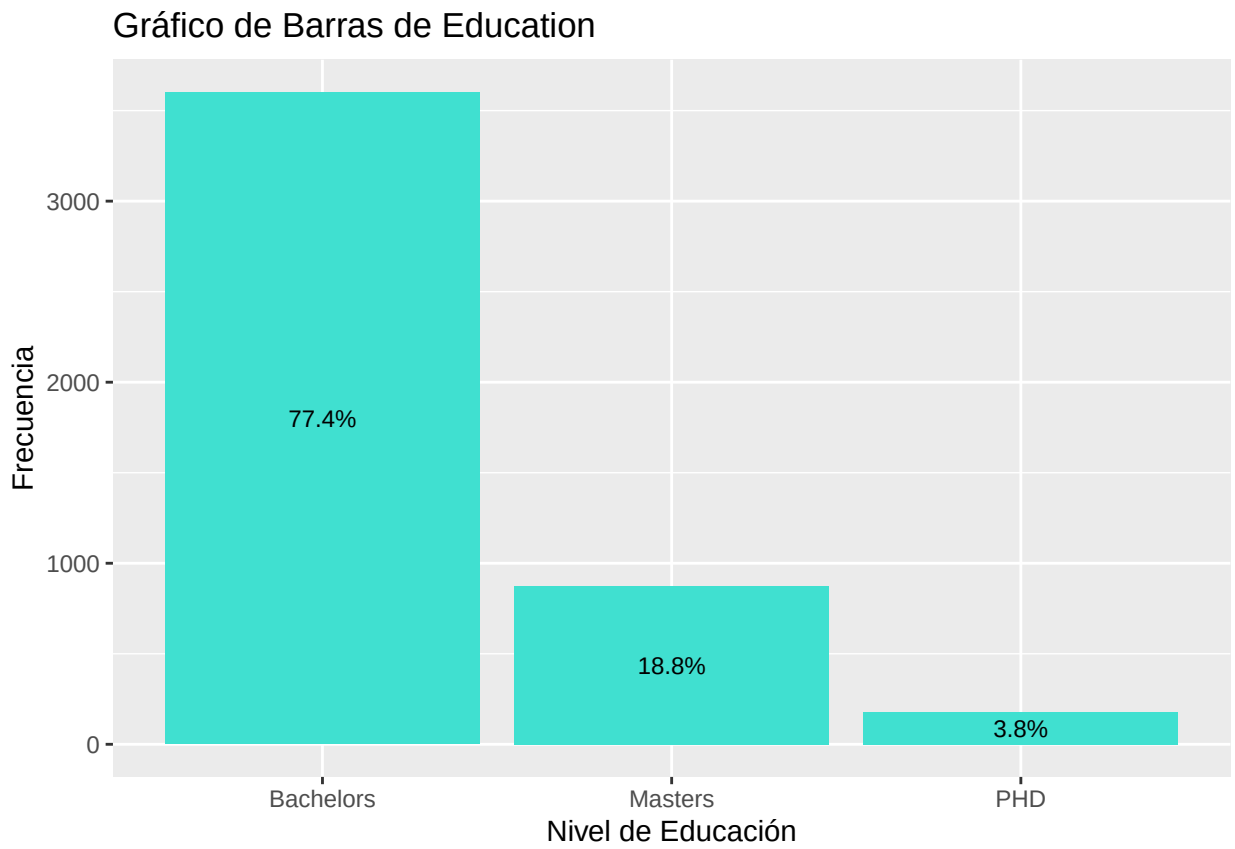


Figure 1: Gráfico de Barras de LeaveOrNot con ggplot2

```
## $Education
## [1] "Bachelors" "Masters"   "PHD"
##
## $City
## [1] "Bangalore" "Pune"      "New Delhi"
##
## $Gender
## [1] "Male"   "Female"
##
## $EverBenched
## [1] "No"  "Yes"
```

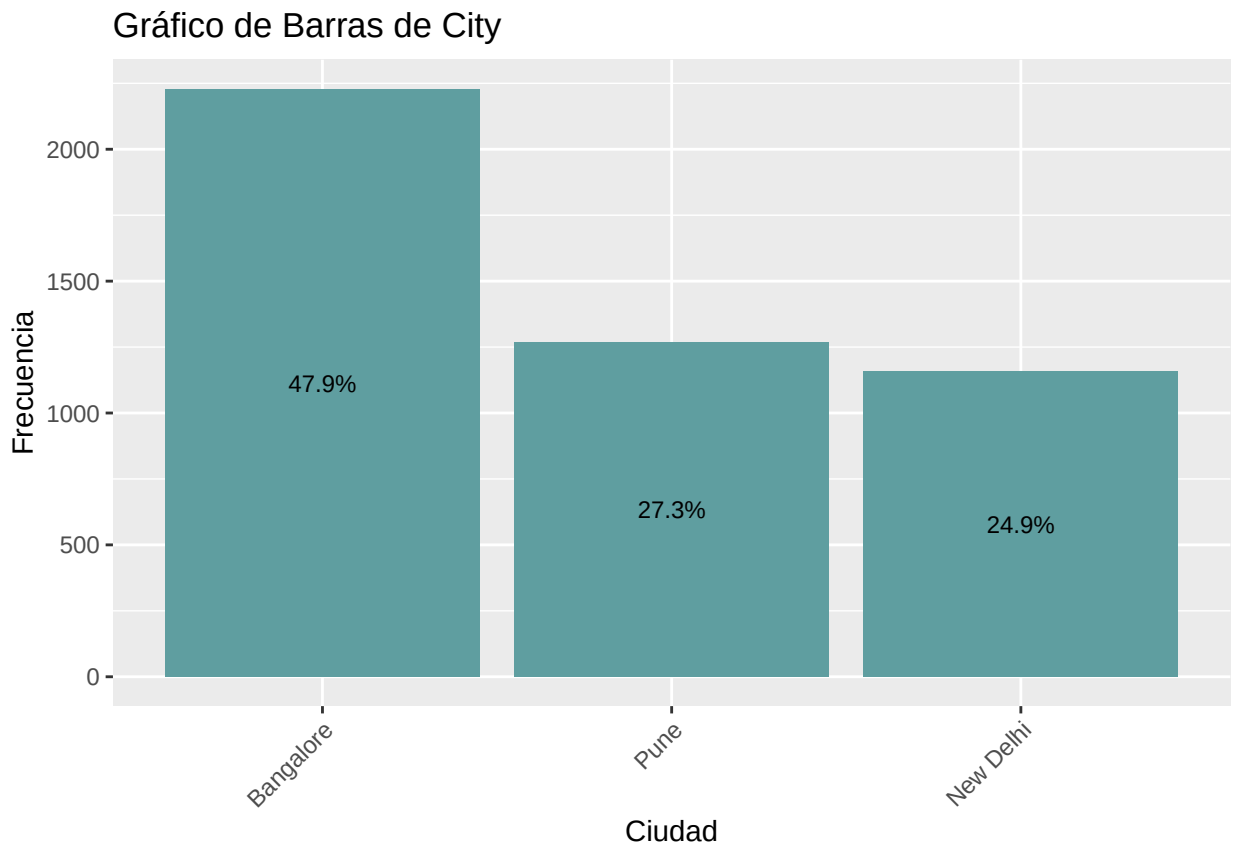
Decidimos realizar un gráfico de barras de Education, debido a que era categórico con pocos valores únicos. Podemos observar que la mayor cantidad de personas tienen un título de Bachelor(77.4%), mientras que una pequeña cantidad de personas tienen un título de PhD (3.8%).

```r
# Calcular los porcentajes de frecuencia de cada categoría de Education
library(ggplot2)
education_data <- as.data.frame(table(datos$Education))
education_data$Percentage <- education_data$Freq / sum(education_data$Freq) * 100

# Crear el gráfico de barras con porcentajes usando ggplot2
ggplot(education_data, aes(x = Var1, y = Freq, label = paste0(round(Percentage, 1), "%"))) +
  geom_bar(stat = "identity", fill = "turquoise") +
  geom_text(size = 3, position = position_stack(vjust = 0.5), color = "black") +
  ggtitle("Gráfico de Barras de Education") +
  xlab("Nivel de Educación") +
  ylab("Frecuencia")
```



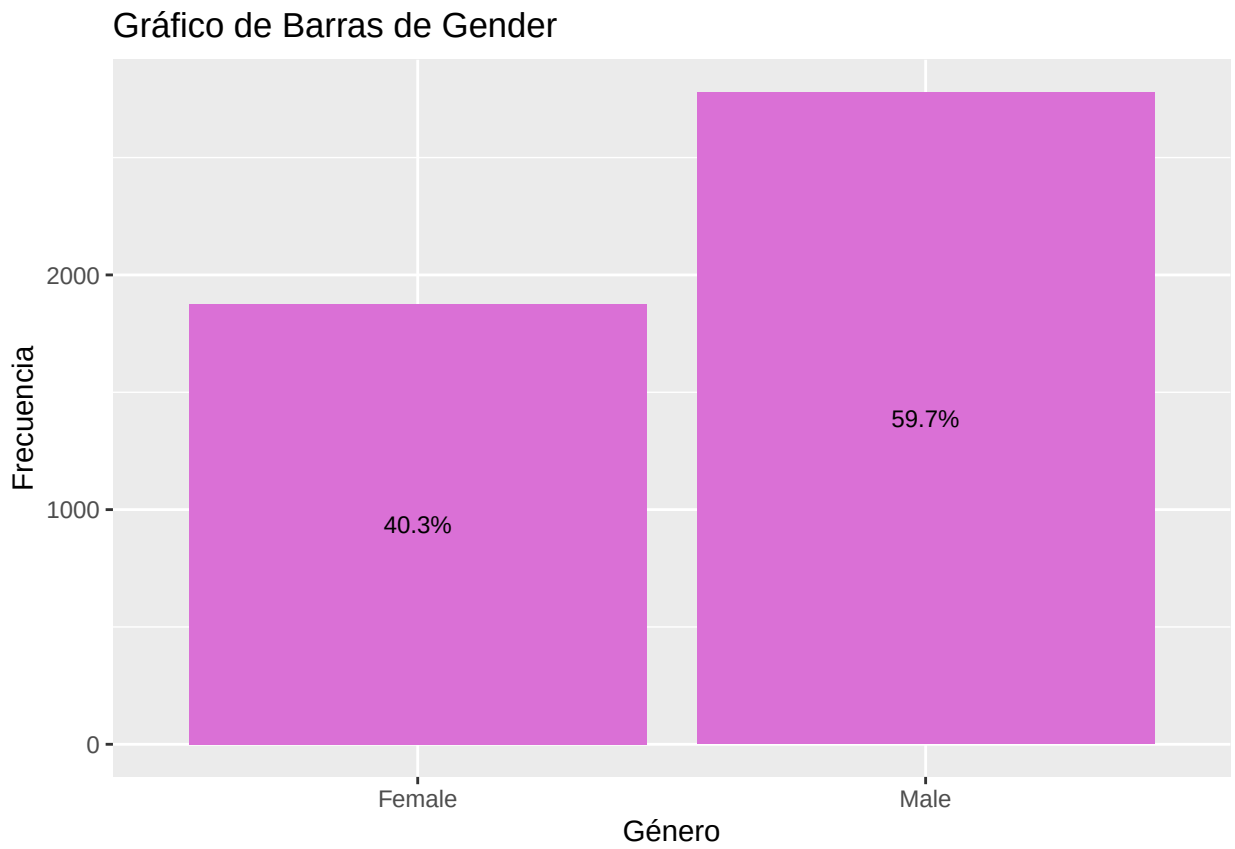
Decidimos realizar un gráfico de barras de City, debido a que era categórico con pocos valores únicos. Podemos observar que la mayor cantidad de personas son de Bangalore (47.9%), mientras que las personas que vienen de Pune y New Delhi tienen proporciones muy similares (27.3% y 24.9% respectivamente).

```
# Calcular los porcentajes de frecuencia de cada ciudad
library(ggplot2)
city_data <- as.data.frame(table(datos$City))
city_data$Percentage <- city_data$Freq / sum(city_data$Freq) * 100

# Crear el gráfico de barras con porcentajes usando ggplot2
ggplot(city_data, aes(x = reorder(Var1, -Freq), y = Freq, label = paste0(round(Percentage, 1), "%"))) +
  geom_bar(stat = "identity", fill = "cadetblue") +
  geom_text(size = 3, position = position_stack(vjust = 0.5), color = "black") +
  ggtitle("Gráfico de Barras de City") +
  xlab("Ciudad") +
  ylab("Frecuencia") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Rotating x-axis labels for better visibili
```



Decidimos realizar un gráfico de barras de Gender, debido a que era categórico con pocos valores únicos. Podemos observar que la mayor cantidad de personas son del género masculino (59.7%), mientras que la menor proporción son del género femenino (40.3%).

```
# Calcular los porcentajes de frecuencia de cada género
library(ggplot2)
gender_data <- as.data.frame(table(datos$Gender))
gender_data$Percentage <- gender_data$Freq / sum(gender_data$Freq) * 100

# Crear el gráfico de barras con porcentajes usando ggplot2
ggplot(gender_data, aes(x = Var1, y = Freq, label = paste0(round(Percentage, 1), "%"))) +
  geom_bar(stat = "identity", fill = "orchid") +
  geom_text(size = 3, position = position_stack(vjust = 0.5), color = "black") +
  ggtitle("Gráfico de Barras de Gender") +
  xlab("Género") +
  ylab("Frecuencia")
```



Decidimos realizar un gráfico de barras de EverBenched, debido a que era categórico con pocos valores únicos. Podemos observar que casi todas las personas han sido benchadas (89.7%), mientras que solo una minoría no lo fue (10.3%).

```
# Calcular los porcentajes de frecuencia de cada categoría de EverBenched
library(ggplot2)
everbenched_data <- as.data.frame(table(datos$EverBenched))
everbenched_data$Percentage <- everbenched_data$Freq / sum(everbenched_data$Freq) * 100

# Crear el gráfico de barras con porcentajes usando ggplot2
ggplot(everbenched_data, aes(x = Var1, y = Freq, label = paste0(round(Percentage, 1), "%"))) +
  geom_bar(stat = "identity", fill = "lightcoral") +
  geom_text(size = 3, position = position_stack(vjust = 0.5), color = "black") +
  ggtitle("Gráfico de Barras de EverBenched") +
  xlab("Ha Sido Bencheado") +
  ylab("Frecuencia")
```

```
# Converting 'LeaveOrNot' to a factor if it's not already
datos$LeaveOrNot <- as.factor(datos$LeaveOrNot)
# Optionally, convert other categorical variables
datos$PaymentTier <- as.factor(datos$PaymentTier)

# Now, subset the dataset with the required variables
selected_variables <- datos[, c("Age", "JoiningYear", "ExperienceInCurrentDomain", "PaymentTier", "Leave
```

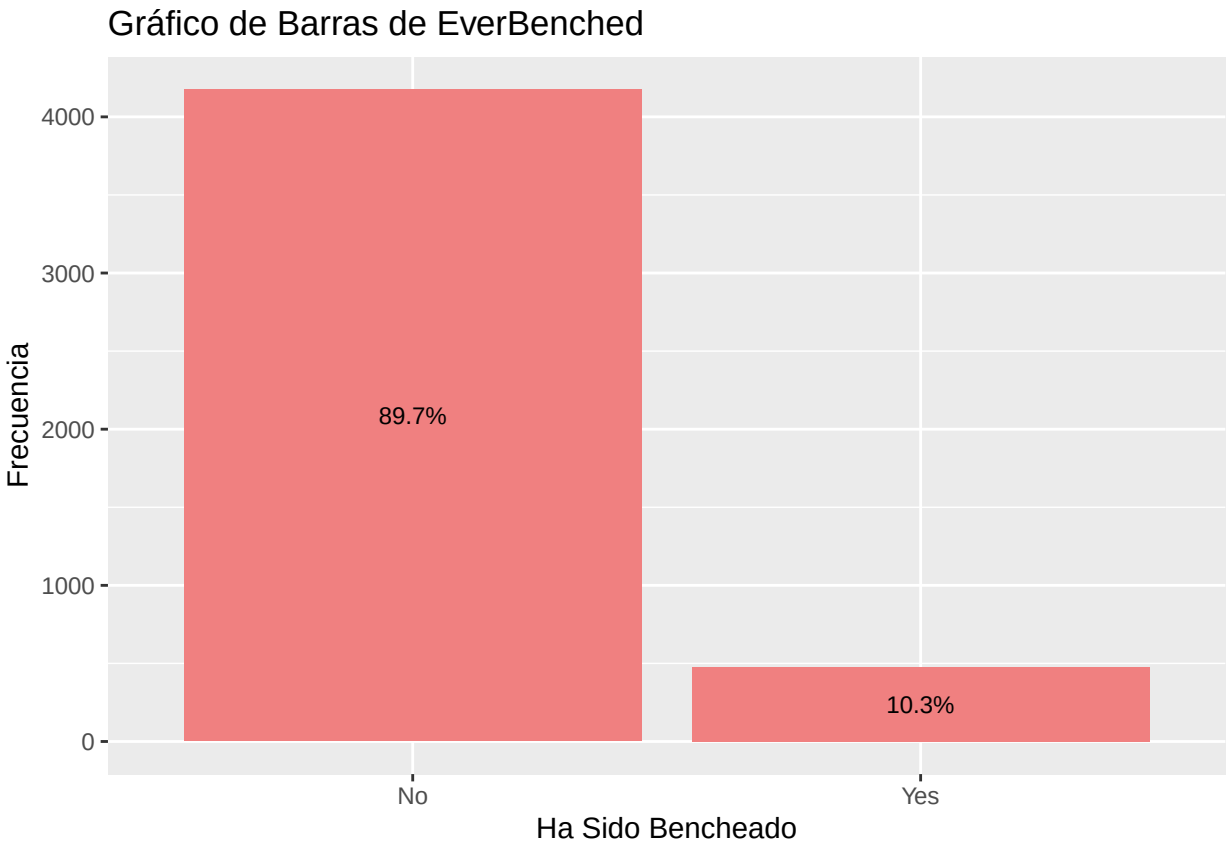


Figure 2: Gráfico de Barras de EverBenched con ggplot2

```
# Creating the pairplot using GGally
library(GGally)
ggpairs(selected_variables, aes(color = LeaveOrNot),
        lower = list(continuous = wrap("points", alpha = 0.5, size = 1)),
        diag = list(continuous = wrap("barDiag")),
        upper = list(continuous = wrap("cor", size = 3)),
        title = "Pairplot with LeaveOrNot as Hue")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

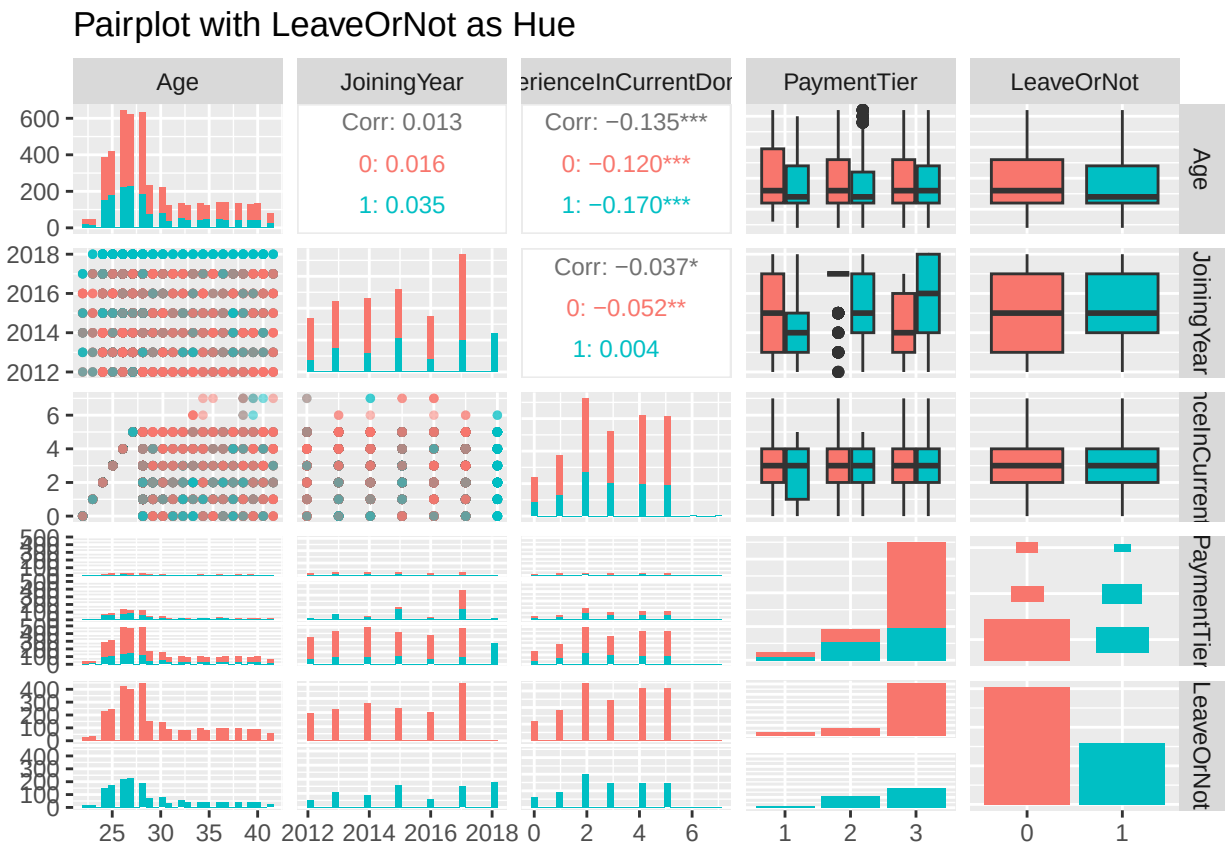


Figure 3: Pairplot with LeaveOrNot as Hue

```
# Subsetting the dataset to include only numerical variables
numerical_data <- datos[sapply(datos, is.numeric)]

# Calculating the correlation matrix
```

```r
cor_matrix <- cor(numerical_data, use="pairwise.complete.obs")  # Handling missing values by considerin

# Display the correlation matrix
cor_matrix
```

```
##                              JoiningYear         Age ExperienceInCurrentDomain
## JoiningYear                   1.00000000  0.01316529               -0.03652462
## Age                           0.01316529  1.00000000               -0.13464285
## ExperienceInCurrentDomain    -0.03652462 -0.13464285                1.00000000
```

```r
# Visualizing the correlation matrix
corrplot(cor_matrix, method="circle", type="upper", order="hclust",
         tl.col="black", tl.srt=45)  # Rotate labels for better readability
```

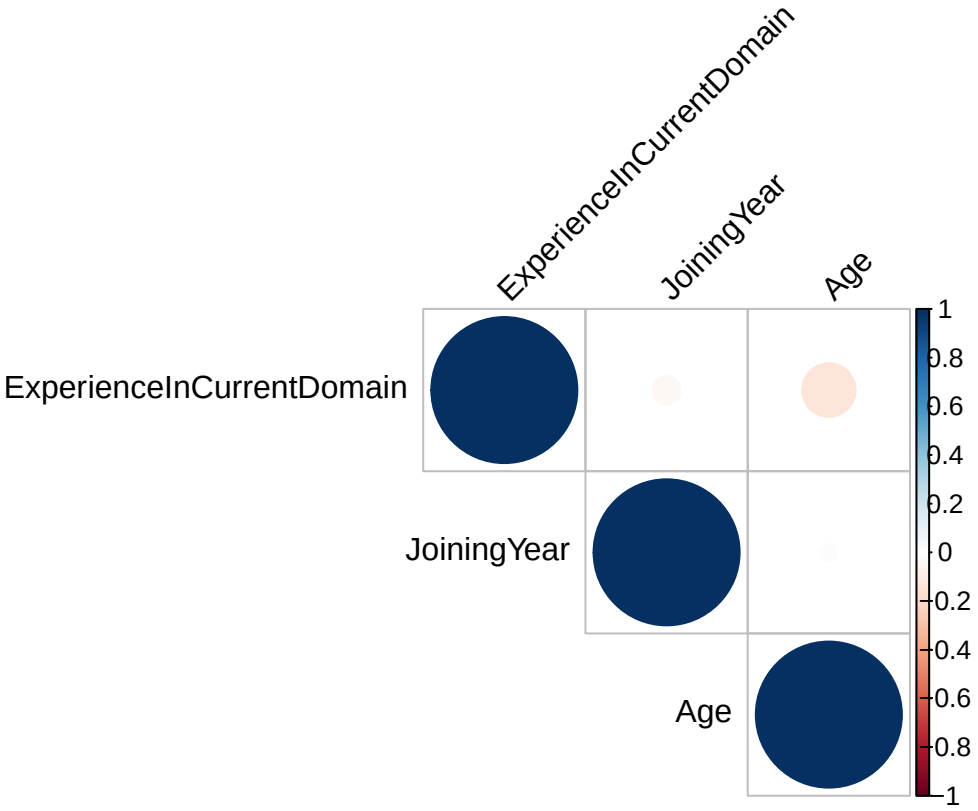


Figure 4: Correlation Matrix Visualization

3. Estadísticos Inferenciales

```r
#install.packages("ggplot2")
#install.packages("conflicted")
```

```r
# Cargamos las librerias necesarias
library(conflicted)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## v readr     2.1.5
```

```r
library(ggplot2)
library(reshape2)
```

```r
# CARGAMOS LOS DATOS
df <- datos
```

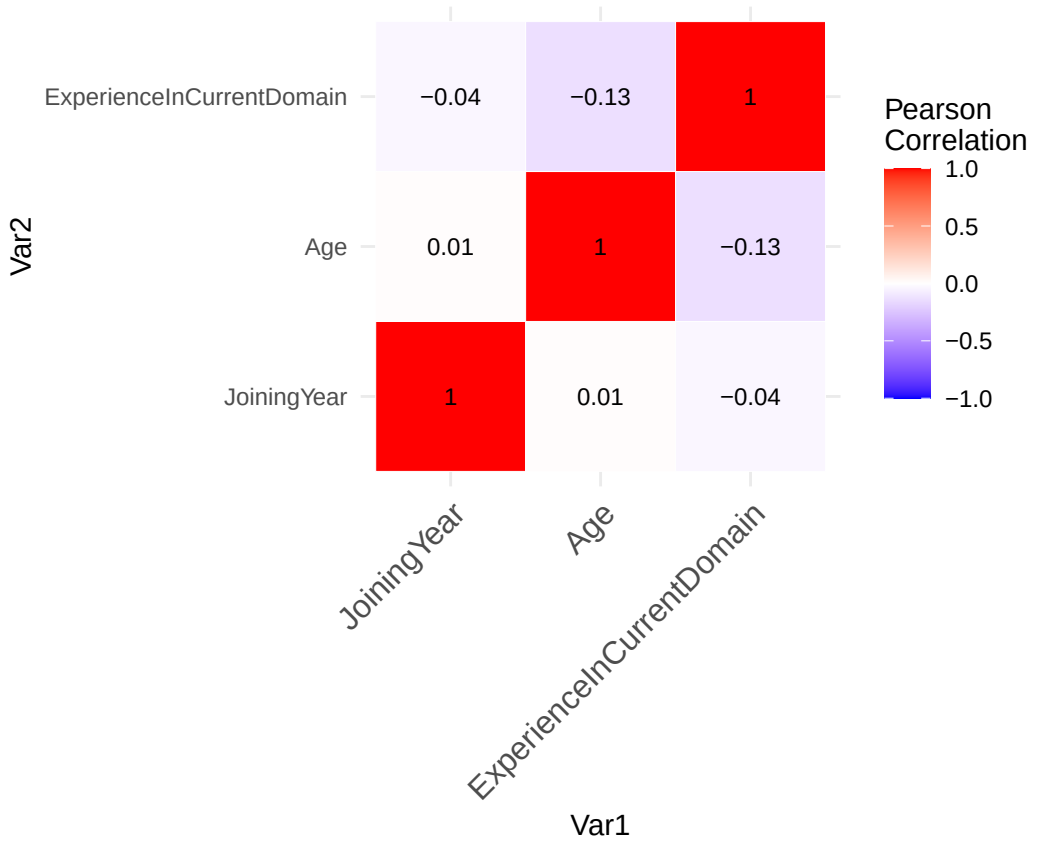
CALCULAMOS LA MATRIZ DE CORRELACION

```r
cor_matrix <- cor(df[, sapply(df, is.numeric)])
cor_matrix
```

```
##                             JoiningYear         Age ExperienceInCurrentDomain
## JoiningYear                  1.00000000  0.01316529                -0.03652462
## Age                          0.01316529  1.00000000                -0.13464285
## ExperienceInCurrentDomain -0.03652462 -0.13464285                 1.00000000
```

VISUALIZAMOS LA MATRIZ DE CORRELACION

```r
hot_matrix <- melt(cor_matrix)
ggplot(hot_matrix, aes(Var1, Var2, fill=value)) +
  geom_tile(color="white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint=0, limit=c(-1,1), space="Lab",
                       name="Pearson\nCorrelation") +
  geom_text(aes(Var1, Var2, label = round(value, 2)), color = "black", size = 3) +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1)) +
  coord_fixed()
```

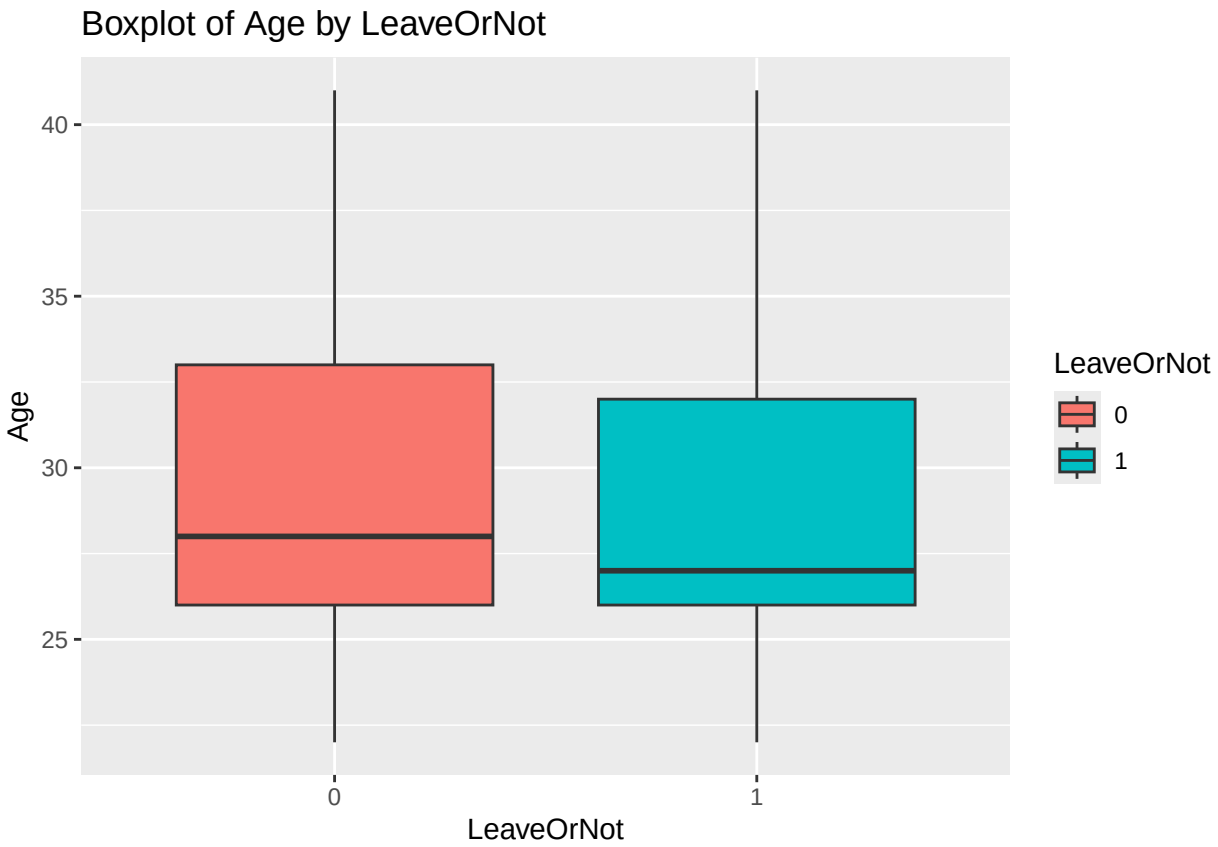


La matriz de correlación muestra las siguientes relaciones entre las variables numéricas:

- La correlación entre las variables no muestra relaciones fuertemente lineales. La mayoría de los coeficientes son cercanos a cero, indicando poca o ninguna correlación lineal.
- La variable LeaveOrNot tiene una correlación positiva de 0.18 con JoiningYear (año de ingreso), lo que podria indicar que mientras más tiempo lleva un empleado en la empresa, es más probable que decida no dejarla.
- La varibale LeaveOrNot tiene una relacion negativa con PaymenTier (nivel de salario) de -0.20, lo que podría indicar que a menor salario un empleado en la emppresa, es mas probable decida dejar la empresa.
- Las otras correlaciones son relativamente bajas, sugiriendo que estas variables no están fuertemente influenciadas entre sí en términos lineales.

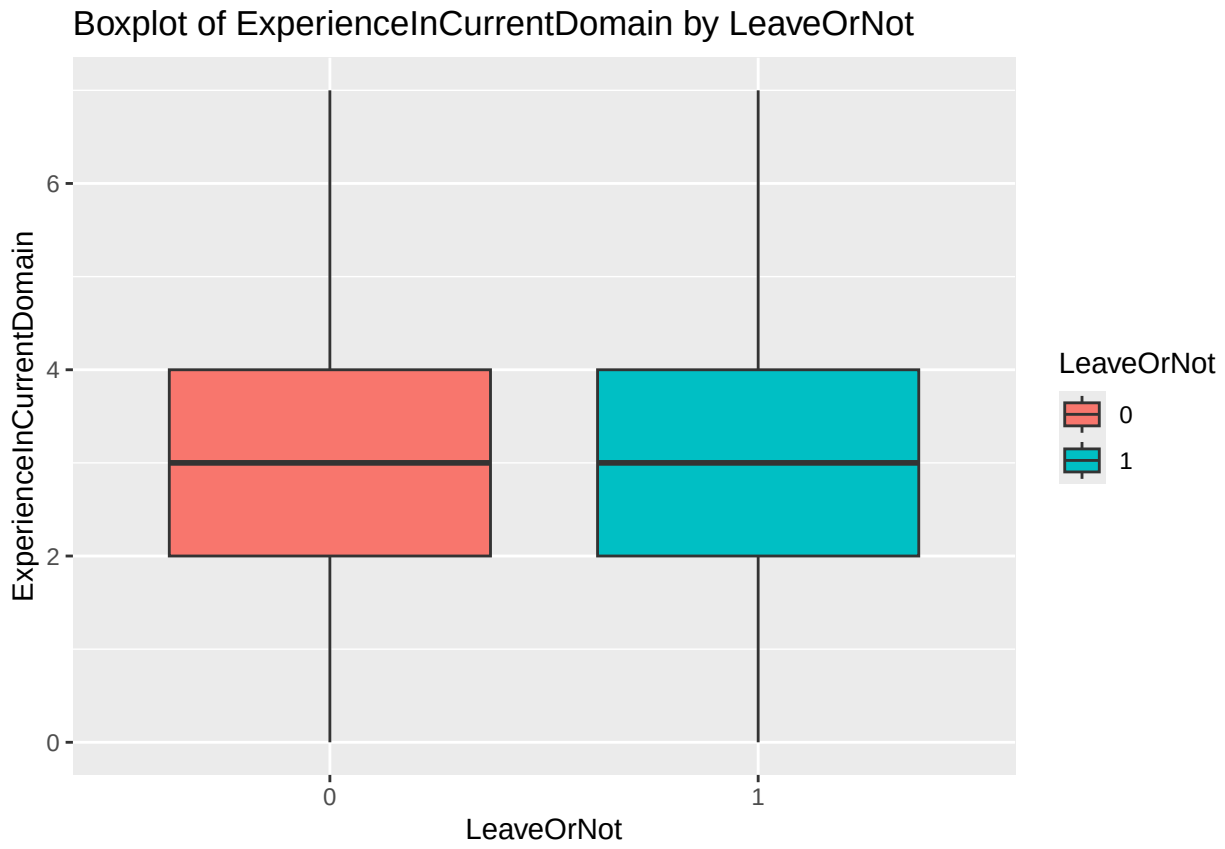
REALIZAMOS UN BOX PLOT PARA VER LAS RELACIONES ENTRE LA VARIABLE EDAD Y LA VARIABLE TARGET

```
ggplot(df, aes(x=LeaveOrNot, y=Age, fill=LeaveOrNot)) +
  geom_boxplot() +
  labs(title = "Boxplot of Age by LeaveOrNot", x = "LeaveOrNot", y = "Age")
```



REALIZAMOS UN BOX PLOT PARA VER LAS RELACIONES ENTRE LA VARIABLE ExperienceInCurrentDomain Y LA VARIABLE TARGET

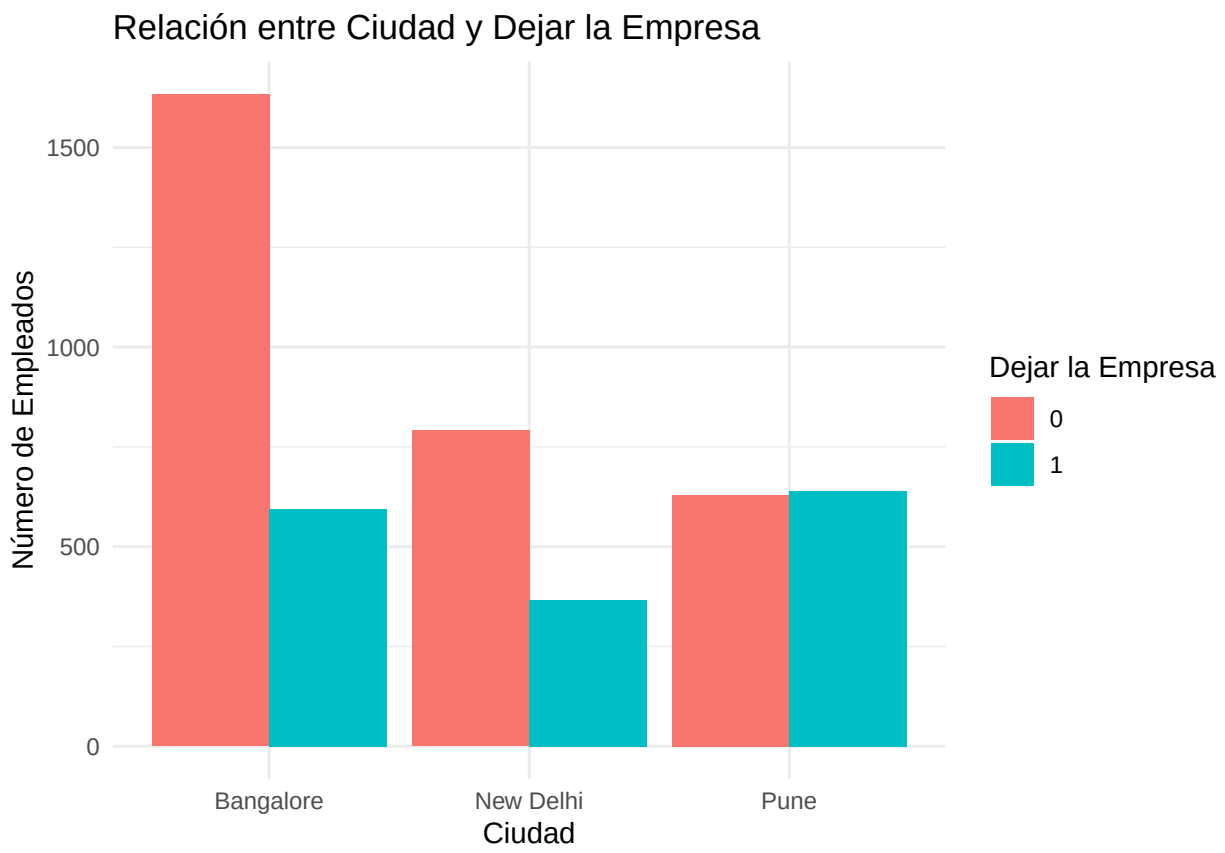
```
ggplot(df, aes(x=LeaveOrNot, y=ExperienceInCurrentDomain, fill=LeaveOrNot)) +
  geom_boxplot() +
  labs(title = "Boxplot of ExperienceInCurrentDomain by LeaveOrNot", x = "LeaveOrNot", y = "ExperienceI
```



- Edad vs. Dejar la Empresa: No parece haber una diferencia significativa en la edad de los empleados que deciden dejar la empresa y los que se quedan.
- Experiencia en el Dominio Actual vs. Dejar la Empresa: Similarmente, la experiencia en el dominio actual no muestra una diferencia clara en términos de afectar la decisión de dejar o no la empresa.

GRAFICO DE RELACION ENTRE VARIABLES "City" VS "LeaveOrNot"

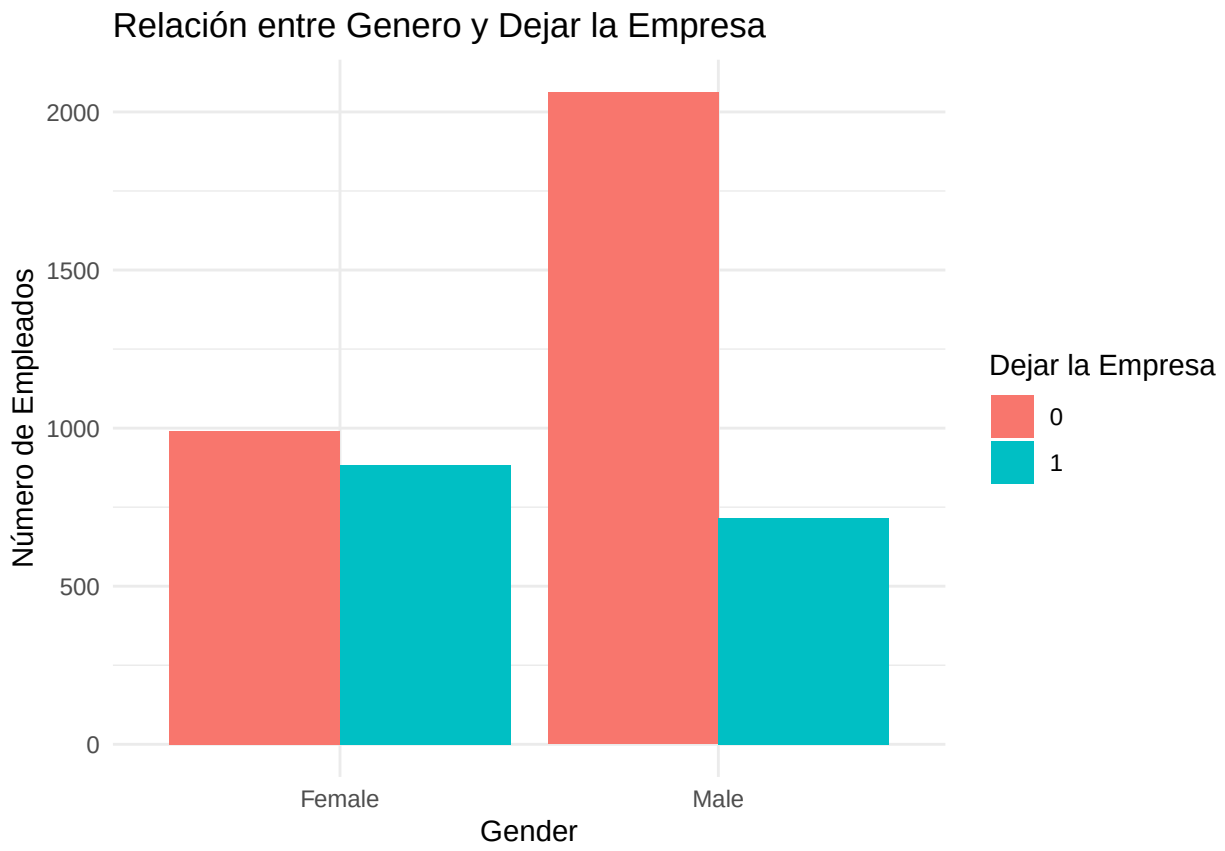
```
ggplot(data = df, aes(x = City, fill = factor(LeaveOrNot))) +  # Define las variables para el eje X y e
  geom_bar(position = "dodge", stat = "count") +  # Crea barras separadas para cada grupo en LeaveOrNot
  labs(title = "Relación entre Ciudad y Dejar la Empresa",
       x = "Ciudad",
       y = "Número de Empleados",
       fill = "Dejar la Empresa") +  # Personaliza las etiquetas y título
  theme_minimal()
```



- Del grafico, se puede observar que la mayoria de los empleados que deciden dejar la empresa son de la ciudad de Bangalore y New Delhi.
- Por otro lado, la ciudad de Pune tiene una cantidad similar de empleados que deciden dejar la empresa y los que se quedan.

GRAFICO DE RELACION ENTRE VARIABLES "Gender" VS "LeaveOrNot"

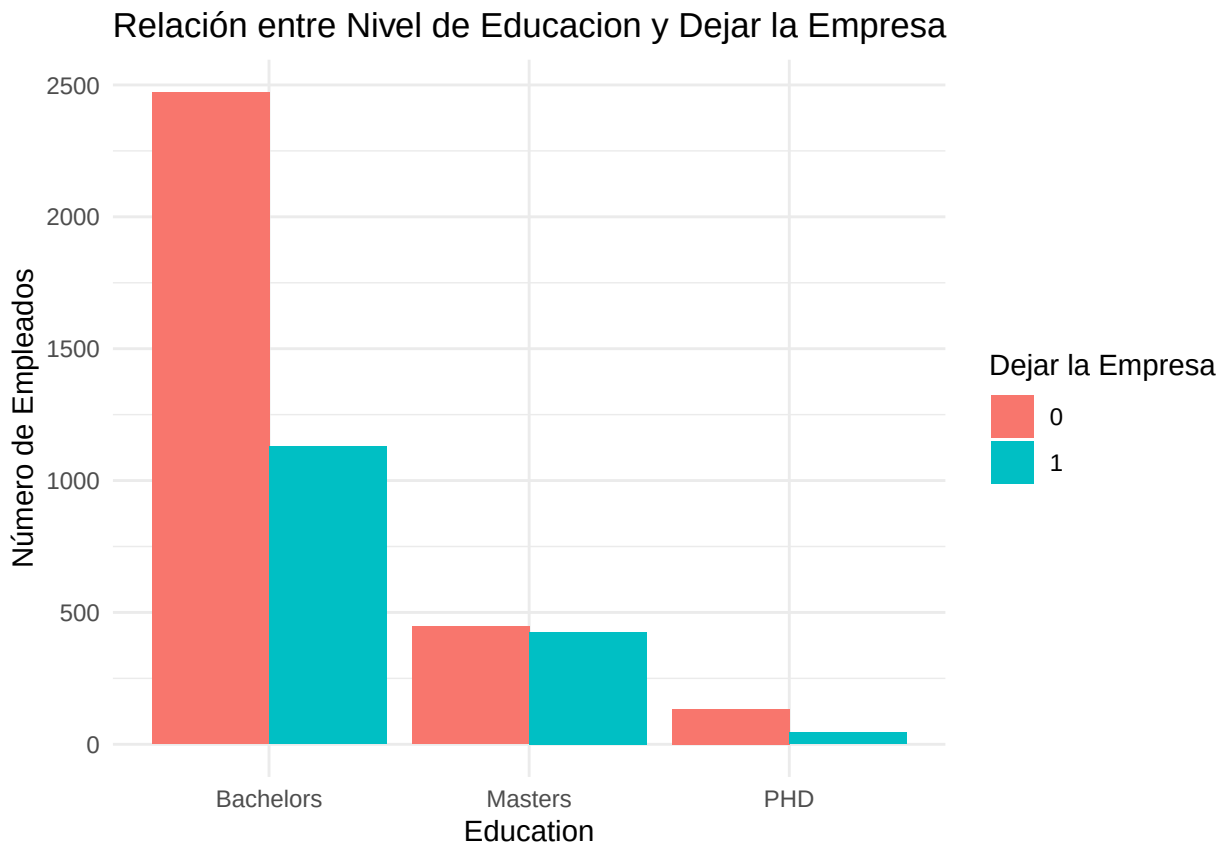
```
ggplot(data = df, aes(x = Gender, fill = factor(LeaveOrNot))) +  # Define las variables para el eje X y
  geom_bar(position = "dodge", stat = "count") +  # Crea barras separadas para cada grupo en LeaveOrNot
  labs(title = "Relación entre Genero y Dejar la Empresa",
       x = "Gender",
       y = "Número de Empleados",
       fill = "Dejar la Empresa") +  # Personaliza las etiquetas y título
  theme_minimal()
```



- Del grafico, se puede observar que la mayoria de los empleados que deciden dejar la empresa son de genero masculino.
- Por otro lado, la cantidad de empleados que deciden quedarse en la empresa es similar para ambos generos.

GRAFICO DE RELACION ENTRE VARIABLES "Education" VS "LeaveOrNot"

```
ggplot(data = df, aes(x = Education, fill = factor(LeaveOrNot))) +  # Define las variables para el eje
  geom_bar(position = "dodge", stat = "count") +  # Crea barras separadas para cada grupo en LeaveOrNot
  labs(title = "Relación entre Nivel de Educacion y Dejar la Empresa",
       x = "Education",
       y = "Número de Empleados",
       fill = "Dejar la Empresa") +  # Personaliza las etiquetas y título
  theme_minimal()
```



- Del grafico, se puede observar que la mayoria de los empleados que deciden dejar la empresa tienen un nivel de educacion de "Bachelors".
- Por otro lado, la cantidad de empleados que deciden quedarse en la empresa es similar para todos los niveles de educacion Master y PHD.

3. Prueba de Hipótesis

Problema:

¿Se puede determinar la renuncia de los empleados en función de las variables independientes del dataset?

-Para determinar si estas diferencias son estadísticamente significativas, procederemos a realizar pruebas de Chi-cuadrado. Estas pruebas nos permitirán evaluar si las diferencias observadas en cada grupo son suficientes para sugerir una dependencia entre las variables categóricas y la decisión de dejar la empresa.

```
# PRUEBA DE CHI-CUADRADO PARA CIUDAD
city_table <- table(df$City, df$LeaveOrNot)
city_table
```

```
##
##                0    1
##   Bangalore 1633  595
##   New Delhi  791  366
##   Pune       629  639
```

```r
CH2_CITY <- chisq.test(city_table)
CH2_CITY
```

```
##
##  Pearson's Chi-squared test
##
## data:  city_table
## X-squared = 206.16, df = 2, p-value < 2.2e-16
```

```r
# PRUEBA DE CHI-CUADRADO PARA GENERO
gender_table <- table(df$Gender, df$LeaveOrNot)
gender_table
```

```
##
##             0    1
##   Female  991  884
##   Male   2062  716
```

```r
CH2_GENDER <- chisq.test(gender_table)
CH2_GENDER
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  gender_table
## X-squared = 225.7, df = 1, p-value < 2.2e-16
```

```r
# PRUEBA DE CHI-CUADRADO PARA EDUCACION
education_table <- table(df$Education, df$LeaveOrNot)
education_table
```

```
##
##                0    1
##   Bachelors 2472 1129
##   Masters    447  426
##   PHD        134   45
```

```r
CH2_EDUCATION <- chisq.test(education_table)
CH2_EDUCATION
```

```
##
##  Pearson's Chi-squared test
##
## data:  education_table
## X-squared = 101.83, df = 2, p-value < 2.2e-16
```

- La prueba de Chi-cuadrado para la variable "City" muestra un valor de p-value $<$ 2.2e-16, lo que indica que hay una relación significativa entre la ciudad de los empleados y su decisión de dejar la empresa.
- La prueba de Chi-cuadrado para la variable "Gender" muestra un valor de p-value $<$ 2.2e-16, lo que indica que hay una relación significativa entre el género de los empleados y su decisión de dejar la empresa.

- La prueba de Chi-cuadrado para la variable “Education” muestra un valor de p-value $< 2.2e\text{-}16$, lo que indica que hay una relación significativa entre el nivel de educación de los empleados y su decisión de dejar la empresa.

CONCLUSIONES

- Los empleados de la ciudad de Bangalore y New Delhi tienen una mayor propensión a dejar la empresa en comparación con los de Pune.
- Los empleados de género masculino tienen una mayor propensión a dejar la empresa en comparación con los de género femenino.
- Los empleados con un nivel de educación de “Bachelors” tienen una mayor propensión a dejar la empresa en comparación con los de “Masters” y “PHD”.