Solución EDA1

Isaac Martín

2024-06-03

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr)

# Cargar el conjunto de datos
url <- "https://archive.ics.uci.edu/ml/machine-learning-databases/00222/bank.zip"
download.file(url, "bank.zip")
unzip("bank.zip", "bank-full.csv")
bank_data <- read.csv("bank-full.csv",sep=";")

# Pregunta 1: Estructura del DataFrame
head(bank_data)
```

```
##   age          job marital education default balance housing loan contact day
## 1  58   management married  tertiary      no    2143     yes   no unknown   5
## 2  44   technician  single secondary      no      29     yes   no unknown   5
## 3  33 entrepreneur married secondary      no       2     yes  yes unknown   5
## 4  47  blue-collar married   unknown      no    1506     yes   no unknown   5
## 5  33      unknown  single   unknown      no       1      no   no unknown   5
## 6  35   management married  tertiary      no     231     yes   no unknown   5
##   month duration campaign pdays previous poutcome  y
## 1   may      261        1    -1        0  unknown no
## 2   may      151        1    -1        0  unknown no
## 3   may       76        1    -1        0  unknown no
## 4   may       92        1    -1        0  unknown no
## 5   may      198        1    -1        0  unknown no
## 6   may      139        1    -1        0  unknown no
```

```
dim(bank_data)
```

```
## [1] 45211    17
```

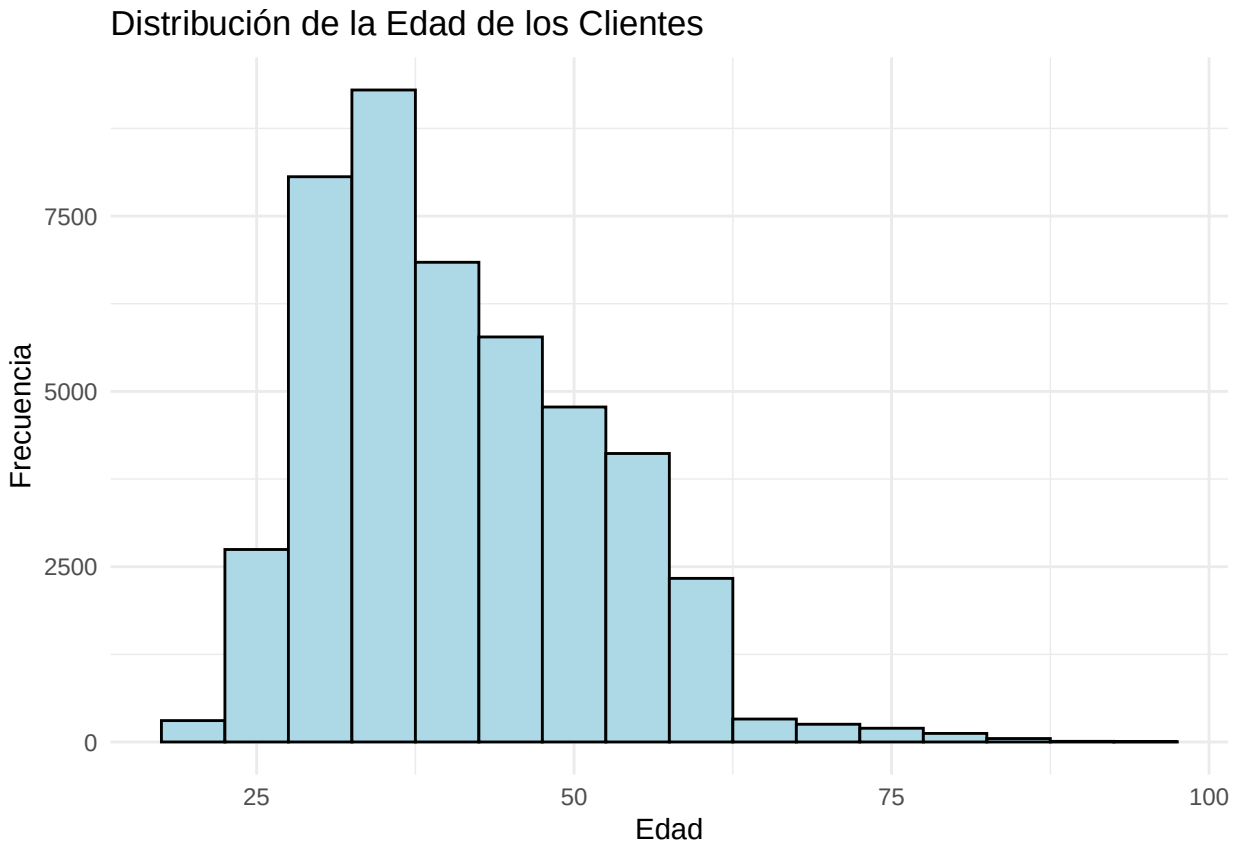
```
# Pregunta 2: Resumen del DataFrame
summary(bank_data)
```

```
##       age            job              marital           education
##  Min.   :18.00   Length:45211       Length:45211       Length:45211
##  1st Qu.:33.00   Class :character   Class :character   Class :character
##  Median :39.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :40.94
##  3rd Qu.:48.00
##  Max.   :95.00
##    default            balance         housing             loan
##  Length:45211       Min.   : -8019   Length:45211       Length:45211
##  Class :character   1st Qu.:    72   Class :character   Class :character
##  Mode  :character   Median :   448   Mode  :character   Mode  :character
##                     Mean   :  1362
##                     3rd Qu.:  1428
##                     Max.   :102127
##    contact               day           month              duration
##  Length:45211       Min.   : 1.00   Length:45211       Min.   :   0.0
##  Class :character   1st Qu.: 8.00   Class :character   1st Qu.: 103.0
##  Mode  :character   Median :16.00   Mode  :character   Median : 180.0
##                     Mean   :15.81                      Mean   : 258.2
##                     3rd Qu.:21.00                      3rd Qu.: 319.0
##                     Max.   :31.00                      Max.   :4918.0
##     campaign          pdays          previous          poutcome
##  Min.   : 1.000   Min.   : -1.0   Min.   :  0.0000   Length:45211
##  1st Qu.: 1.000   1st Qu.: -1.0   1st Qu.:  0.0000   Class :character
##  Median : 2.000   Median : -1.0   Median :  0.0000   Mode  :character
##  Mean   : 2.764   Mean   : 40.2   Mean   :  0.5803
##  3rd Qu.: 3.000   3rd Qu.: -1.0   3rd Qu.:  0.0000
##  Max.   :63.000   Max.   :871.0   Max.   :275.0000
##       y
##  Length:45211
##  Class :character
##  Mode  :character
##
##
##
```

```r
# Pregunta 3: Distribución de la Edad
summary(bank_data$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   33.00   39.00   40.94   48.00   95.00
```

```r
ggplot(bank_data, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "black") +
  labs(title = "Distribución de la Edad de los Clientes",
       x = "Edad",
       y = "Frecuencia") +
  theme_minimal()
```




```
# Pregunta 4: Balance Promedio
mean(bank_data$balance, na.rm = TRUE)
```

```
## [1] 1362.272
```

```
mean(bank_data$balance[bank_data$y == "yes"], na.rm = TRUE)
```

```
## [1] 1804.268
```

```
# Pregunta 5: Frecuencia de Contacto
summary(bank_data$campaign)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   2.764   3.000  63.000
```

```
# Pregunta 6: Análisis de Duración
summary(bank_data$duration)
```

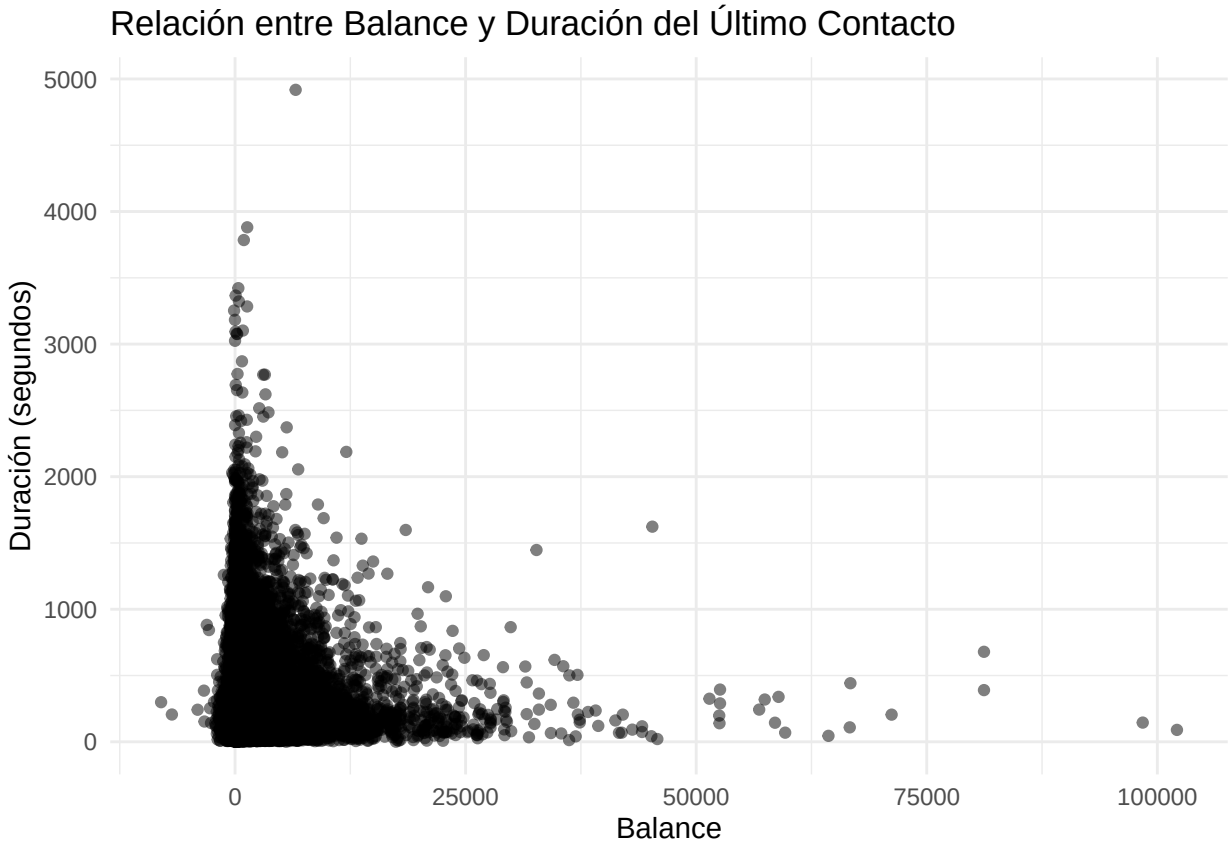
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   103.0   180.0   258.2   319.0  4918.0
```

```
mean(bank_data$duration[bank_data$y == "yes"], na.rm = TRUE)
```

```
## [1] 537.2946
```

```
# Pregunta 7: Relación entre Balance y Duración
ggplot(bank_data, aes(x = balance, y = duration)) +
  geom_point(alpha = 0.5) +
  labs(title = "Relación entre Balance y Duración del Último Contacto",
       x = "Balance",
       y = "Duración (segundos)") +
```

```
  theme_minimal()
```




```
# Pregunta 8: Segmentación por Trabajo
bank_data %>% group_by(job) %>%
  summarise(media_balance = mean(balance, na.rm = TRUE),
            mediana_balance = median(balance, na.rm = TRUE))
```

```
## # A tibble: 12 x 3
##    job           media_balance mediana_balance
##    <chr>                 <dbl>           <dbl>
##  1 admin.                1136.             396
##  2 blue-collar           1079.             388
##  3 entrepreneur          1521.             352
##  4 housemaid             1392.             406
##  5 management            1764.             572
##  6 retired               1984.             787
##  7 self-employed         1648.             526
##  8 services               997.             340.
##  9 student               1388.             502
## 10 technician            1253.             421
## 11 unemployed            1522.             529
## 12 unknown               1772.             677
```

```
# Pregunta 9: Análisis de Contactos Anteriores
summary(bank_data$pdays)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    -1.0    -1.0    -1.0    40.2    -1.0   871.0
```

```r
mean(bank_data$pdays[bank_data$y == "yes"], na.rm = TRUE)
```

```
## [1] 68.70297
```

```r
# Pregunta 10: Estudio General
# Puedes utilizar técnicas de análisis más avanzadas como correlación, regresión, etc.
cor(bank_data$balance, bank_data$duration, use = "complete.obs")
```

```
## [1] 0.02156038
```

```r
ggplot(bank_data, aes(x = campaign, y = balance)) +
  geom_point(alpha = 0.5) +
  labs(title = "Relación entre Campañas y Balance",
       x = "Número de Campañas",
       y = "Balance") +
  theme_minimal()
```