# Aplicación de conceptos de R

Tatiana Casallas Marin

18/08/2025

## Tabla de contenido



install_tinytex() install.packages(“tidyverse”)

# 1 Contexto

Un concesionario desea perfilar a sus clientes con el fin de mejorar sus estrategias en ventas. Para ello, ha creado una tabla que contiene información sobre aspectos socieconómicos de sus clientes.
El Objetivo de esta actividad consiste en practicar el uso de herramientas del lenguaje de programación R para analizar la base de datos denominada *base_concesionario*, y generar un informe que de cuenta de los hallazgos, metodologías y exploración de los datos.

# 2 Actividades

## 2.1 Análisis exploratorio de la base de datos

a. Cargar la base de datos denominada *base_concesionario*.

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.5.1
```

```
base_concesionario <- read_excel("base_concesionario.xlsx")
View(base_concesionario)
```

b. Realizar una exploración de la base de tal manera que se pueda definir cuántos clientes están registrados, qué variables están asociadas a los clientes.

```r
summary(base_concesionario)
```

```
##    PERSONA              EDAD              SEXO             ESTATURA
##  Length:62          Length:62          Length:62          Length:62
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  NIVEL ESCOLAR      MARCA DE AUTO      NUMERO DE HIJOS       SALARIO
##  Length:62          Length:62          Length:62          Min.   : 800000
##  Class :character   Class :character   Class :character   1st Qu.:2000000
##  Mode  :character   Mode  :character   Mode  :character   Median :3450000
##                                                           Mean   :3286667
##                                                           3rd Qu.:4700000
##                                                           Max.   :6500000
##                                                           NA's   :2
##    MASCOTA
##  Length:62
##  Class :character
##  Mode  :character
##
##
##
##
```

```r
table(base_concesionario$`NIVEL ESCOLAR`)
```

```
##
##   DOCTORADO    MAESTRÍA          NA         PhD PROFESIONAL
##          14          20           1           4          20
```

Tenemos en cuenta las variables persona, edad, sexo, estatura, nivel escolar, marca de auto, numero de hijos, salario y si tienen mascotas, hay en total 62 datos, aunque de estos hay 2 filas de datos faltantes, con la funcion table dice cuantas datos hay por variable

c. Identificar y contar los datos faltantes que se tengan en la base de datos.

```r
d<- is.na.data.frame(base_concesionario)

summary(d)
```

```
##   PERSONA           EDAD            SEXO           ESTATURA
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:60        FALSE:60        FALSE:59        FALSE:60
##  TRUE :2         TRUE :2         TRUE :3         TRUE :2
##  NIVEL ESCOLAR   MARCA DE AUTO   NUMERO DE HIJOS  SALARIO
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:59        FALSE:58        FALSE:59        FALSE:60
##  TRUE :3         TRUE :4         TRUE :3         TRUE :2
```

```
##   MASCOTA
##  Mode :logical
##  FALSE:59
##  TRUE :3
```

la cantidad de datos que son iguales a TRUE son la cantidad de datos faltantes, que en total son 16

d. Utilizar y aplicar estrategias que permitan identificar registros incompletos, valores inconsistentes y otras características que podrían afectar el análisis de los datos de la base.

```
na<- colSums(is.na(base_concesionario))
print(na)
```

```
##         PERSONA            EDAD            SEXO        ESTATURA   NIVEL ESCOLAR
##               2               2               3               2               3
##   MARCA DE AUTO NUMERO DE HIJOS         SALARIO         MASCOTA
##               4               3               2               3
```

```
filas_con_na <- base_concesionario[!complete.cases(base_concesionario), ]
print("Filas con valores NA:")
```

```
## [1] "Filas con valores NA:"
```

```
print(filas_con_na)
```

```
## # A tibble: 8 x 9
##   PERSONA EDAD  SEXO  ESTATURA `NIVEL ESCOLAR` `MARCA DE AUTO` `NUMERO DE HIJOS`
##   <chr>   <chr> <chr> <chr>    <chr>           <chr>           <chr>
## 1 <NA>    <NA>  <NA>  <NA>     <NA>            <NA>            <NA>
## 2 <NA>    <NA>  <NA>  <NA>     <NA>            <NA>            <NA>
## 3 PERSON~ 68    F     1.65     MAESTRÍA        <NA>            2
## 4 PERSON~ 60    F     1.63     MAESTRÍA        FORD            <NA>
## 5 PERSON~ NA    F     1.54     <NA>            FORD            NA
## 6 PERSON~ 21    <NA>  3.45     NA              BMW             0
## 7 PERSON~ 20    M     1.78     MAESTRÍA        CHEVROLET       0
## 8 PERSON~ 68    F     1.65     PROFESIONAL     <NA>            3
## # i 2 more variables: SALARIO <dbl>, MASCOTA <chr>
```

## 2.2 Análisis de la variable “marca de auto”

a. Evaluar la variable “Marca de auto” y determinar si hay datos faltantes, en caso de que los haya, se requiere proponer una metodología que permita tratar con este tipo de datos.

```
sum(is.na(base_concesionario$`MARCA DE AUTO`))
```

```
## [1] 4
```

```r
base_concesionario$`MARCA DE AUTO`[is.na(base_concesionario$`MARCA DE AUTO`)] <- "DESCONOCIDO"

table(base_concesionario$`MARCA DE AUTO`)
```

```
##
##        AUDI         BMW         BWM   CHEVROLET DESCONOCIDO         FOR
##          13          11           1          12           4           1
##        FORD          NA     renault     RENAULT
##           6           1           1          12
```

usamos el sum para ver la cantidad total de datos que son iguales a Na, lo que hicimos fue ca,biar los datos Na por la variable desconocidos con eso la funcion table sale mas completa

b. Crear una tabla de frecuencias que permita determinar la preferencia en marcas de autos de los clientes.

```r
tabla1<- table(base_concesionario$`MARCA DE AUTO`)

print(tabla1)
```

```
##
##        AUDI         BMW         BWM   CHEVROLET DESCONOCIDO         FOR
##          13          11           1          12           4           1
##        FORD          NA     renault     RENAULT
##           6           1           1          12
```

d. Haciendo uso de la librería *ggplot2* de R, generar un gráfico de barras y de sectores que permita visualizar la distribución de las marcas de autos entre los clientes.

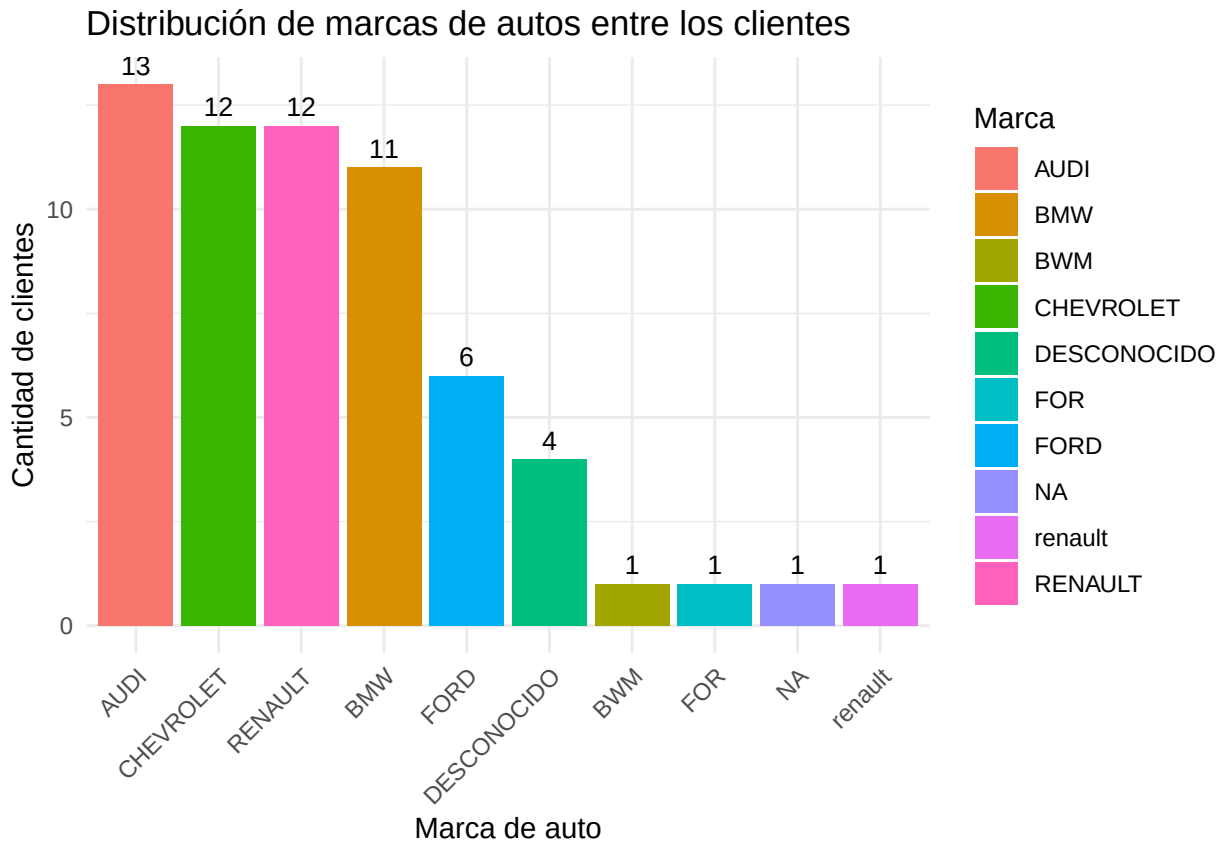
```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.5.1
```

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.5.1

##
## Adjuntando el paquete: 'dplyr'

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

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

```r
conteo_marcas <- base_concesionario %>%
  count(`MARCA DE AUTO`) %>%
  rename(cantidad = n) %>%
  mutate(porcentaje = round(cantidad / sum(cantidad) * 100, 1))

conteo_marcas <- conteo_marcas %>%
  filter(!is.na(`MARCA DE AUTO`))

grafico_barras <- ggplot(conteo_marcas, aes(x = reorder(`MARCA DE AUTO`, -cantidad), y = cantidad, fill
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(
    title = "Distribución de marcas de autos entre los clientes",
    x = "Marca de auto",
    y = "Cantidad de clientes",
    fill = "Marca"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  geom_text(aes(label = cantidad), vjust = -0.5, size = 3.5)

print(grafico_barras)
```




e. Proporcionar una conclusión a partir de la información que se haya extraído en el desarrollo de los numerales anteriores.

1). Con el grafico vemos que hay datos repetidos pero estos al estar mal escritos hacen que el analisis se vuelva complejo, por lo que es fundamentl que la informacion sea clara

## 2.3 Análisis de la variable edad

a. Verificar si la variable edad está correctamente definida como tipo *numérico*.

```
class(base_concesionario$EDAD)
```

```
## [1] "character"
```

b. Identificar algún tipo de anomalía en la variable y si es el caso corregirla.

```
base_concesionario$EDAD <- as.numeric(base_concesionario$EDAD)
```

```
## Warning: NAs introducidos por coerción
```

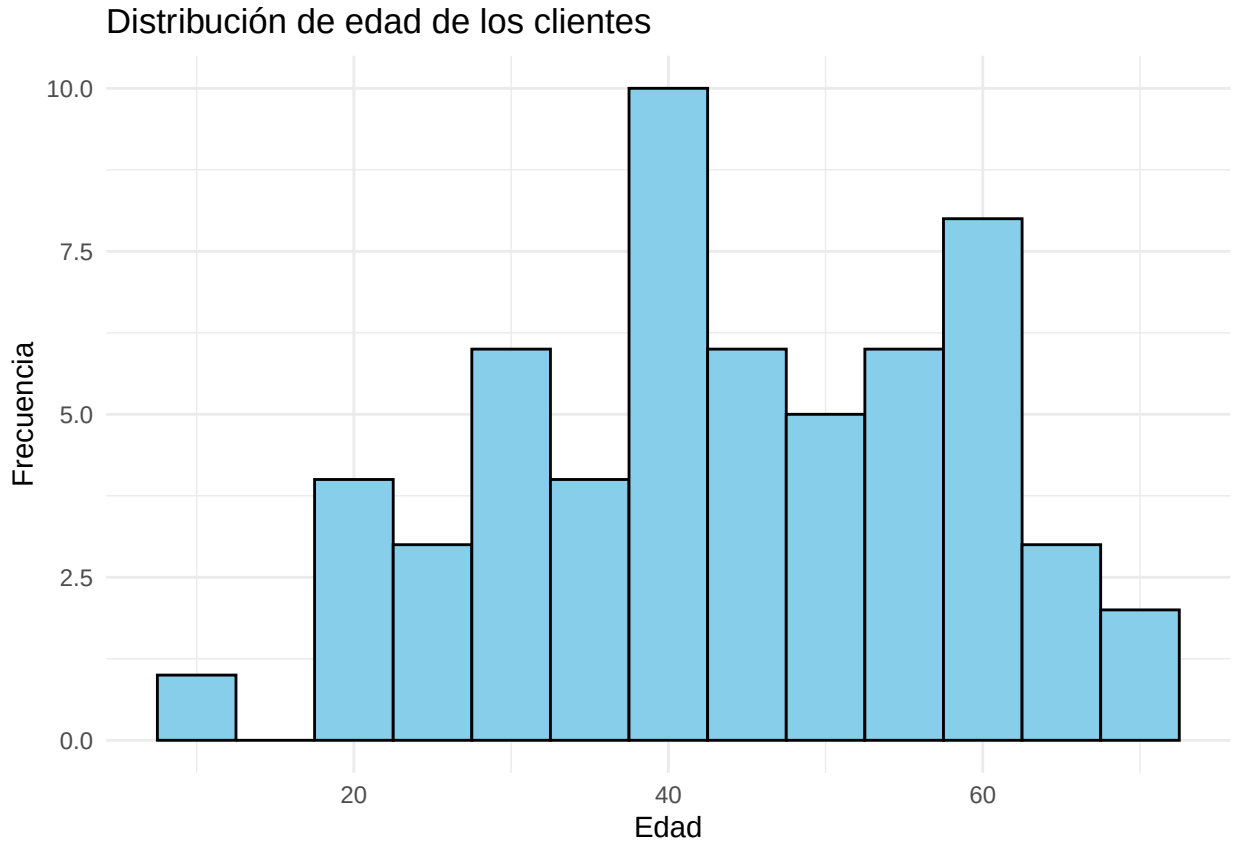
```
class(base_concesionario$EDAD)
```

```
## [1] "numeric"
```

c. Haciendo uso de la librería *ggplot2* de R, realizar un histogramas de la variable edad y describir características de distribución de la variable.

```
histograma_edad <- ggplot(base_concesionario, aes(x = EDAD)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  labs(
    title = "Distribución de edad de los clientes",
    x = "Edad",
    y = "Frecuencia"
  ) +
  theme_minimal()

histograma_edad
```

```
## Warning: Removed 4 rows containing non-finite outside the scale range
## (`stat_bin()`).
```


Distribución de edad de los clientes
Frecuencia
10.0
7.5
5.0
2.5
0.0
20
40
60
Edad