Taller1

Victor Julio Pe{a

2025-07-26

# Actividades

## Análisis exploratorio de la base de datos

### a. Cargar la base de datos denominada *base\_concesionario*.

library(readxl)  
library(dplyr)

##   
## 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

concesionario <- read\_excel("C:/Users/HP/Downloads/base\_concesionario.xlsx")  
concesionario <- concesionario[-c(1,2),] #se vizualizo la base y hay dos filas sin registros por lo tanto se eliminan  
head(concesionario, 4)

## # A tibble: 4 × 9  
## PERSONA EDAD SEXO ESTATURA `NIVEL ESCOLAR` `MARCA DE AUTO` `NUMERO DE HIJOS`  
## <chr> <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 PERSON… 21 M 1.54 MAESTRÍA AUDI 0   
## 2 PERSON… 26 F 1.55 PROFESIONAL RENAULT 5   
## 3 PERSON… 30 F 1.6 DOCTORADO BMW 2   
## 4 PERSON… 31 f 1.7 PROFESIONAL RENAULT 2   
## # ℹ 2 more variables: SALARIO <dbl>, MASCOTA <chr>

### 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.

dim(concesionario) # Filas y columnas (cuántos clientes y variables)

## [1] 60 9

str(concesionario) # Estructura de los datos

## tibble [60 × 9] (S3: tbl\_df/tbl/data.frame)  
## $ PERSONA : chr [1:60] "PERSONA 1" "PERSONA 2" "PERSONA 3" "PERSONA 4" ...  
## $ EDAD : chr [1:60] "21" "26" "30" "31" ...  
## $ SEXO : chr [1:60] "M" "F" "F" "f" ...  
## $ ESTATURA : chr [1:60] "1.54" "1.55" "1.6" "1.7" ...  
## $ NIVEL ESCOLAR : chr [1:60] "MAESTRÍA" "PROFESIONAL" "DOCTORADO" "PROFESIONAL" ...  
## $ MARCA DE AUTO : chr [1:60] "AUDI" "RENAULT" "BMW" "RENAULT" ...  
## $ NUMERO DE HIJOS: chr [1:60] "0" "5" "2" "2" ...  
## $ SALARIO : num [1:60] 1200000 1250000 900000 800000 950000 2000000 2500000 3500000 4700000 5000000 ...  
## $ MASCOTA : chr [1:60] "SI" "NO" "NO" "NO" ...

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

concesionario[concesionario == "NA"] <- NA  
colSums(is.na(concesionario))

## PERSONA EDAD SEXO ESTATURA NIVEL ESCOLAR   
## 0 1 1 0 2   
## MARCA DE AUTO NUMERO DE HIJOS SALARIO MASCOTA   
## 3 2 0 1

### 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.

#valores inconsistentes  
unique(concesionario$SEXO)

## [1] "M" "F" "f" "MUJER" NA "HOMBRE" "m" "mujer"   
## [9] "hombre"

# Convertir todo a minúsculas   
concesionario$SEXO <- tolower(concesionario$SEXO)  
  
# Normalizar valores comunes  
concesionario$SEXO[concesionario$SEXO %in% c("m", "masculino", "h", "hombre", "male")] <- "M"  
concesionario$SEXO[concesionario$SEXO %in% c("f", "femenino", "fem", "mujer", "female")] <- "F"  
  
# Verificar que solo queden las dos categorías  
unique(concesionario$SEXO)

## [1] "M" "F" NA

# 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.

# Revisar datos faltantes en la variable  
sum(is.na(concesionario$`MARCA DE AUTO`))

## [1] 3

# eliminar filas con na  
base\_limpia <- concesionario %>% filter(!is.na(`MARCA DE AUTO`))

Se elimanron las filas con datos NA en la variable marca de auto

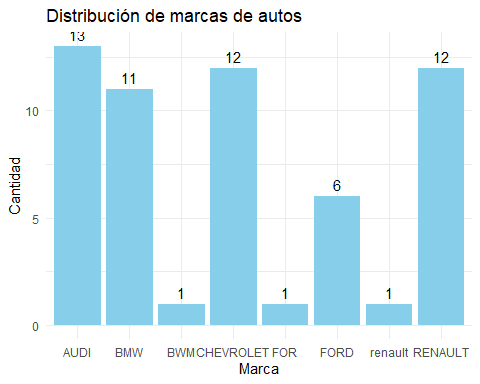
### b. Crear una tabla de frecuencias que permita determinar la preferencia en marcas de autos de los clientes.

frecuencia <- table(base\_limpia$`MARCA DE AUTO`)  
porcentajes <- round(prop.table(frecuencia)\*100, 2)  
tabla <- cbind(Frecuencia = frecuencia, Porcentaje = porcentajes)  
print(tabla)

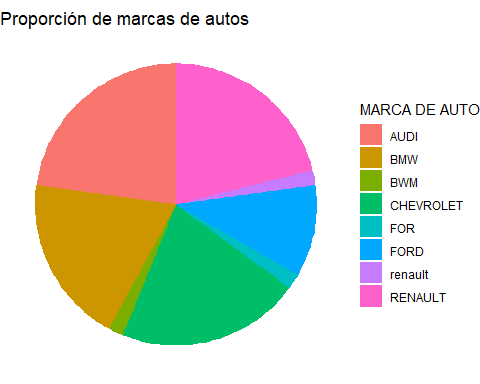
## Frecuencia Porcentaje  
## AUDI 13 22.81  
## BMW 11 19.30  
## BWM 1 1.75  
## CHEVROLET 12 21.05  
## FOR 1 1.75  
## FORD 6 10.53  
## renault 1 1.75  
## RENAULT 12 21.05

### c. 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

library(ggplot2)  
  
freq\_table <- base\_limpia %>%  
 count(`MARCA DE AUTO`)  
  
ggplot(freq\_table, aes(x = `MARCA DE AUTO`, y = n)) +  
 geom\_bar(stat = "identity", fill = "skyblue") +  
 geom\_text(aes(label = n), vjust = -0.5) + # Ajusta para que el texto quede encima  
 theme\_minimal() +  
 labs(title = "Distribución de marcas de autos",  
 x = "Marca",  
 y = "Cantidad")



# Gráfico de sectores (pastel)  
base\_limpia %>%   
 count(`MARCA DE AUTO`) %>%  
 ggplot(aes(x = "", y = n, fill = `MARCA DE AUTO`)) +  
 geom\_col(width = 1) +  
 coord\_polar(theta = "y") +  
 theme\_void() +  
 labs(title = "Proporción de marcas de autos")



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

El análisis de la variable “Marca de auto” reveló una distribución clara en las preferencias de los clientes del concesionario. hay una marca (AUDI) con 13, que predomina preo no por mucho, ya que hay 4 marcas mas que estan por debajo de este no por mucho los cuales son (CHEVROLET Y RENAULT) con 12 y (BMW) con 11.

# 2. Análisis de la variable Edad

### a. Verificar que la variable Edad sea numérica

if (!is.numeric(base\_limpia$EDAD)) {  
 base\_limpia$EDAD <- as.numeric(as.character(base\_limpia$EDAD))  
}

## Warning: NAs introducidos por coerción

class(base\_limpia$EDAD)

## [1] "numeric"

### b. Identificar anomalías y corregirlas

summary(base\_limpia$EDAD)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 10.00 32.50 42.00 43.25 56.00 65.00 2

# Detectar edades negativas o poco probables (>100 años) (<15 años)  
anomalías <- base\_limpia$EDAD[base\_limpia$EDAD < 15 | base\_limpia$EDAD > 90]  
print(anomalías)

## [1] NA NA 10

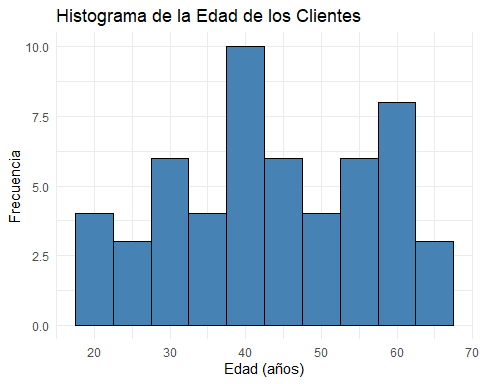
hay una persona menor a 10 años

# eliminar filas con na  
base\_limpia <- base\_limpia %>% filter(!is.na(`EDAD`))  
# Eliminar filas con Edad menor a 15 años  
base\_limpia <- base\_limpia[base\_limpia$EDAD >= 15 | is.na(base\_limpia$EDAD), ]

se eliminan filas con NA en la variable edad y la persona menor a 15 años

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

library(ggplot2)  
  
ggplot(base\_limpia, aes(x = EDAD)) +  
 geom\_histogram(binwidth = 5, fill = "steelblue", color = "black") +  
 theme\_minimal() +  
 labs(title = "Histograma de la Edad de los Clientes",  
 x = "Edad (años)",  
 y = "Frecuencia")



# Análisis de la variable sexo

### a. Revisar la consistencia de la variable “SEXO” y realizar correcciones en caso de encontrar valores inusuales o inconsistentes.

# Ver valores únicos actuales  
unique(base\_limpia$SEXO)

## [1] "M" "F" NA

Anteriormente se hizo una limpieza, por lo tanto no hay valores inconsistentes

base\_limpia <- base\_limpia[!is.na(base\_limpia$SEXO), ]

se eliminan datos con NA

### b. Crear una tabla de frecuencias y extraer conclusiones sobre la proporción de hombres y de mujeres en el conjunto de clientes.

library(dplyr)  
base\_limpia %>%  
 filter(!is.na(SEXO)) %>%  
 count(SEXO) %>%  
 mutate(porcentaje = n / sum(n) \* 100)

## # A tibble: 2 × 3  
## SEXO n porcentaje  
## <chr> <int> <dbl>  
## 1 F 25 47.2  
## 2 M 28 52.8

Tras limpiar la variable SEXO, se encontró que la proporción de hombres y mujeres en la base es aproximadamente 52.8% hombres y 47.2% mujeres. La variable está bien definida y puede usarse para análisis posteriores de segmentación o comportamiento según género.