ANALISIS EXPLORATORIO 1 CONCESIONARIO DE AUTOS

Alejandra Cifuentes / Giovanny Porras

2025-05-24

Índice

2 Exploracion de datos

a. Descargar el archivo TABLA_TALLER.xlsx
b. Cargar el archivo de datos en RStudio

Rta: Carga de datos inicial

```
#datos_base <- read_excel("D:/MaestriaAnalitica/BasesAnalitica/gitRepository/Proyecto-02-Autos/BASE/t1f
datos_base <- read_excel("C:/Users/PC/Documents/ANALITICA/analiticsO1Cars/BASE/t1fe-tabla_taller.xlsx")
kable(head(datos_base, 10,  caption = "Datos iniciales"), format = "latex", booktabs = TRUE) %>%
  kable_styling(latex_options = c("scale_down", "hold_position"))
```

PERSONA	EDAD	SEXO	ESTATURA	NIVEL ESCOLAR	MARCA DE AUTO	NUMERO DE HIJOS	SALARIO	MASCOTA
NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA
PERSONA 1	21	M	1.54	MAESTRÍA	AUDI	0	1200000	SI
PERSONA 2	26	F	1.55	PROFESIONAL	RENAULT	5	1250000	NO
PERSONA 3	30	F	1.6	DOCTORADO	BMW	2	900000	NO
PERSONA 4	31	f	1.7	PROFESIONAL	RENAULT	2	800000	NO
PERSONA 5	35	M	1.71	MAESTRÍA	AUDI	1	950000	NO
PERSONA 6	65	M	1.8	MAESTRÍA	AUDI	1	2000000	SI
PERSONA 7	45	M	1.54	MAESTRÍA	BMW	1	2500000	NO
PERSONA 8	42	F	1.52	PROFESIONAL	RENAULT	1	3500000	SI

c. Describir brevemente la estructura del conjunto de datos: ¿Cuantos clientes estan registrados y que variables incluyen?

```r
no_datos_persona <- datos_base %>%
  filter(grepl("PERSONA", PERSONA)) %>%
  count()

cabeceras_datos <- names(datos_base)
```

Rta: El conjunto de datos tiene 60 clientes registrados e incluyen las varibles PERSONA, EDAD, SEXO, ESTATURA, NIVEL ESCOLAR, MARCA DE AUTO, NUMERO DE HIJOS, SALARIO, MASCOTA .

d. Realizar una exploracion rapida utilizando funciones como head(), tail(), str(), summary(), **Rta:** Los ultimos valores son (tail)

```r
ultimos_datos <- tail(datos_base)
kable(head(ultimos_datos, 10, caption = "Datos ultima posicion"), format = "latex", booktabs = TRUE) %>%
```

PERSONA	EDAD	SEXO	ESTATURA	NIVEL ESCOLAR	MARCA DE AUTO	NUMERO DE HIJOS	SALARIO	MASCOTA
PERSONA 55	30	F	1.54	MAESTRÍA	CHEVROLET	2	2400000	SI
PERSONA 56	39	M	1.58	MAESTRÍA	AUDI	1	2600000	NO
PERSONA 57	34	F	1.6	DOCTORADO	BMW	1	3500000	SI
PERSONA 58	24	f	1.7	PROFESIONAL	RENAULT	3	800000	SI
PERSONA 59	20	M	1.71	MAESTRÍA	AUDI	0	850000	NO
PERSONA 60	10	M	1.8	PROFESIONAL	AUDI	0	1000000	NO

Funciones adicionales:

```r
#print(str(datos_base))
#print(dim(datos_base))
#print(colnames(datos_base))
print(summary(datos_base))
```

```
##    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
##
##
##
##
```

e. Identificar si hay datos faltantes y cuantificar cuantos son en total y por variable

```
#is.na(datos_base)
total_na = datos_base %>% is.na %>% sum()
#Contar NA por columna(variable)
na_por_columna <- colSums(is.na(datos_base))
tabla_na <- data.frame(
   Variable = names(na_por_columna),
   total_na = as.vector(na_por_columna)
)

#Contar NA total
#rowSums(is.na(datos_base))

#Impresion de tabla
#kable(tabla_na, caption = "Variables faltantes por cuantificar")  %>%
#  kable_styling(full_width = FALSE, position = "left")
```

Rta: - El numero de total de datos faltantes es **24** y por variable son los siguientes:

Table 1: Variables faltantes por cuantificar

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

- Analisis: Hay problemas de datos en todas las variables sera importante discriminar cada caso

f. Comentar sobre los posibles problemas en los datos:

Filas vacías al inicio del archivo: - Las dos primeras filas están vacías o no contienen datos válidos.

Inconsistencias categóricas:

- Variabilidad en la codificación de la variable SEXO, con valores como f, mujer, hombre, nan o minúsculas inconsistentes.
- Formatos no estandarizados en NIVEL ESCOLAR, como el uso de PhD en lugar de DOCTORADO.
- Uso de minúsculas en valores de MARCA DE AUTO, como renault.

Valores faltantes:

- Algunas filas tienen múltiples variables vacías, como en el caso de la PERSONA 24.

Valores extremos o anómalos (outliers):

- PERSONA 31: valor de ESTATURA = 3.45 m, fuera del rango fisiológico normal.
- PERSONA 33: valor de NUMERO DE HIJOS = 54, ampliamente fuera del promedio observado (3.03).

3 Análisis de la variable Marca "Marca de auto".

a. Evaluar la variable **"MARCA DE AUTO"** y determinar si hay faltantes.

```
#datos_base$`MARCA DE AUTO`(is.na)
#(is.na(datos_base$`MARCA DE AUTO`))
data_na_marca_autos <- sum(is.na(datos_base$`MARCA DE AUTO`))
```

Rta: Se tratan datos faltantes y se encuentran un total de **4**; se realiza tratamiento de datos remplazando los "na"/"NA" a "NO_MANEJA", se convierte todo mayúsculas.

```
Sys.setlocale("LC_ALL", "Spanish_Colombia.UTF-8")
```

```
## [1] "LC_COLLATE=Spanish_Colombia.utf8;LC_CTYPE=Spanish_Colombia.utf8;LC_MONETARY=Spanish_Colombia.uti
```

```
#Eliminar columna 1, 2
datos_base <- datos_base[-c(1,2), ]
#Datos Eliminados
#print(datos_base)
na_marca_autos <- sum(is.na(datos_base$`MARCA DE AUTO`))
# PERSONA 13, 32, 49, NA y datos vacios
datos_base$`MARCA DE AUTO`[is.na(datos_base$`MARCA DE AUTO`) | datos_base$`MARCA DE AUTO` == "NA"] <- 
# PERSONA 39 cambia FOR A FORD
datos_base$`MARCA DE AUTO`[ datos_base$`MARCA DE AUTO` == "FOR"] <- "FORD"
# PERSONA 40 cambia BWM A BMW
datos_base$`MARCA DE AUTO`[ datos_base$`MARCA DE AUTO` == "BWM"] <- "BMW"

#print(datos_base)
# PERSONA 36 cambio a mayuscula marca
datos_base$`MARCA DE AUTO` <- toupper(datos_base$`MARCA DE AUTO`)
print(datos_base)
```

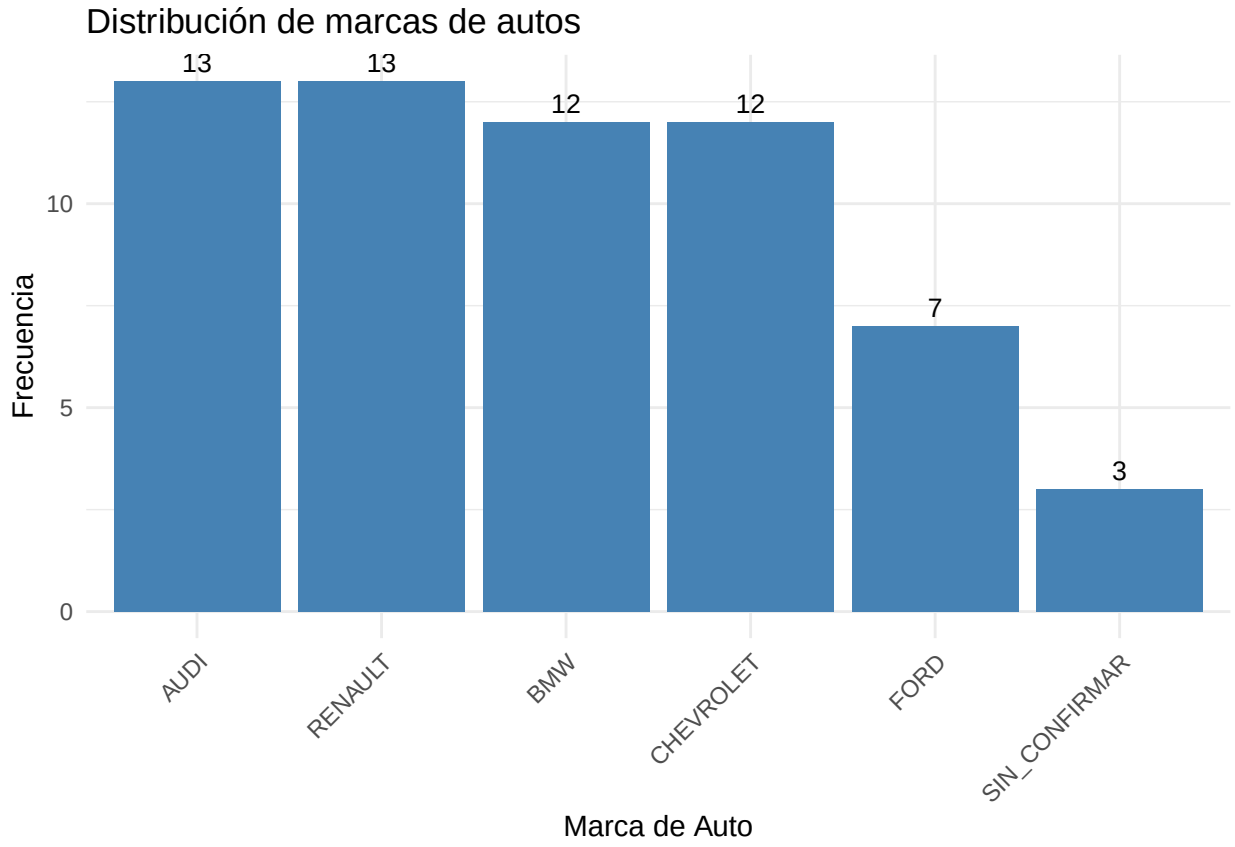
```
## # A tibble: 60 x 9
##    PERSONA    EDAD  SEXO  ESTATURA `NIVEL ESCOLAR` `MARCA DE AUTO`
##    <chr>      <chr> <chr> <chr>    <chr>           <chr>
##  1 PERSONA 1  21    M     1.54     MAESTRÍA        AUDI
##  2 PERSONA 2  26    F     1.55     PROFESIONAL     RENAULT
##  3 PERSONA 3  30    F     1.6      DOCTORADO       BMW
##  4 PERSONA 4  31    f     1.7      PROFESIONAL     RENAULT
##  5 PERSONA 5  35    M     1.71     MAESTRÍA        AUDI
##  6 PERSONA 6  65    M     1.8      MAESTRÍA        AUDI
##  7 PERSONA 7  45    M     1.54     MAESTRÍA        BMW
##  8 PERSONA 8  42    F     1.52     PROFESIONAL     RENAULT
##  9 PERSONA 9  52    F     1.51     DOCTORADO       RENAULT
## 10 PERSONA 10 63    M     1.65     DOCTORADO       RENAULT
## # i 50 more rows
## # i 3 more variables: `NUMERO DE HIJOS` <chr>, SALARIO <dbl>, MASCOTA <chr>
```

b. Crear un tabla de frecuencias para entender la popularidad de las diferentes marcas entre los cliente. **Rta:** En la tabla de frecuencias se observa que las marcas más populares son AUDI, CHEVROLET, RENAULT y BMW. Los tres usuarios que no tienen la categoría " "SIN_CONFIRMAR" " se identificaron como personas a partir de 50 años

```r
tabla_frecuencias_autos <- as.data.frame(table(datos_base$`MARCA DE AUTO`))
names(tabla_frecuencias_autos) <- c("MARCA DE AUTO", "Frecuencia")
#Organizar Mayor a menor
tabla_frecuencias_autos <- tabla_frecuencias_autos %>% arrange(desc(Frecuencia))
print(tabla_frecuencias_autos)
```

```
##   MARCA DE AUTO Frecuencia
## 1          AUDI         13
## 2       RENAULT         13
## 3           BMW         12
## 4     CHEVROLET         12
## 5          FORD          7
## 6 SIN_CONFIRMAR          3
```

c. Generar graficos de barras y de tortas para visualizar la distribucion de las marcas de autos y proporcionar una interpretacion. **Rta:** Se identifica que las marcas de vehiculos preferidas por los clientes del concesionario son **AUDi** y **RENAULT**; se observa una segmentacion igual por pares entre las entre AUDI/RENAULT y CHEVROLET/BMW



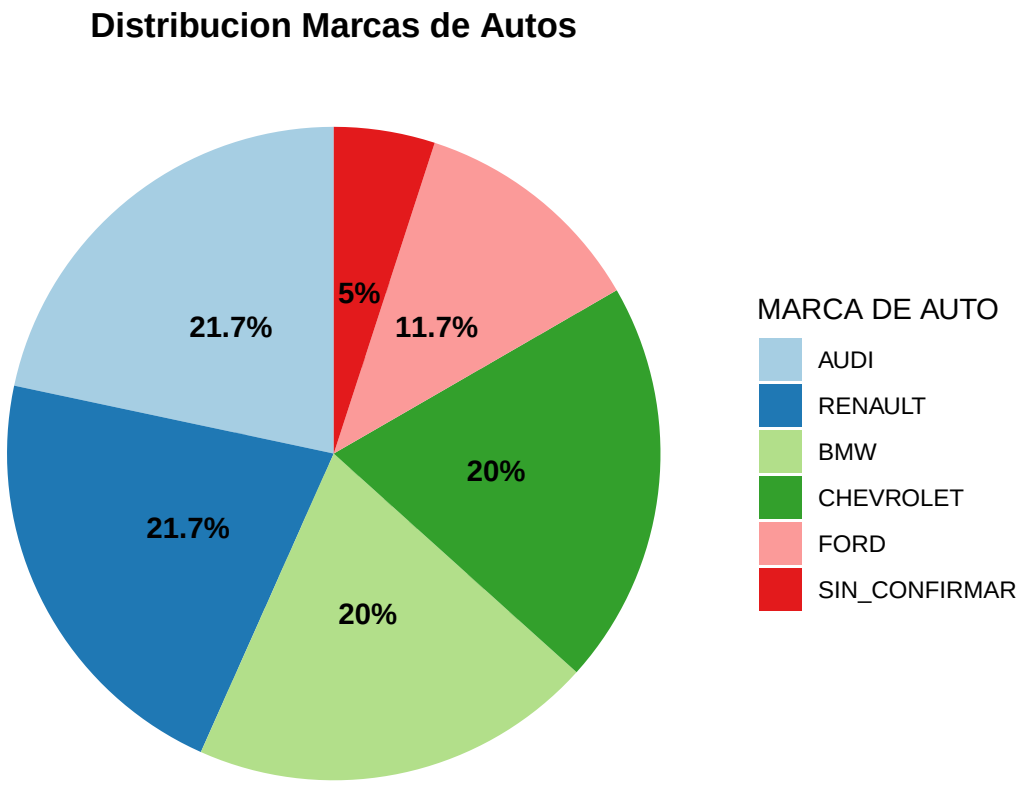
Rta: La preferencia de vehículos esta distribuida entre 4 marcas que suman 83% de la muestra, estas cuatro marcas a su vez representan dos segmentos que se reparten en el mismo porcentaje 21%/21% y 20%/20%.

d. Concluir cuál es la marca de auto más popular entre los clientes. **Rta:** Se concluyen que AUDI y RENAULT son las marcas lideres con una distribución uniforme en el los clientes del concesionario en un porcentaje del 21.7 entre ambas ocupan el 42.4%.

```r
tabla_frecuencias_autos$Porcentaje <- round(tabla_frecuencias_autos$Frecuencia/sum(tabla_frecuencias_au
print(tabla_frecuencias_autos)
```

```
##   MARCA DE AUTO Frecuencia Porcentaje
## 1          AUDI         13       21.7
## 2       RENAULT         13       21.7
## 3           BMW         12       20.0
## 4     CHEVROLET         12       20.0
## 5          FORD          7       11.7
## 6 SIN_CONFIRMAR          3        5.0
```

```r
tabla_frecuencias_autos$`MARCA DE AUTO` <- factor(
  tabla_frecuencias_autos$`MARCA DE AUTO`,
  levels = tabla_frecuencias_autos$`MARCA DE AUTO`[order(-tabla_frecuencias_autos$Frecuencia)]
)
# Crear el grafico circular de MARCAS AUTOS
ggplot(tabla_frecuencias_autos, aes(x = "", y = Frecuencia, fill = `MARCA DE AUTO`)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  geom_text(aes(label = paste0(Porcentaje, "%")),
            position = position_stack(vjust = 0.5),
            size = 4,
            color = "black",
            fontface = "bold") +
  labs(title = "Distribucion Marcas de Autos",
       fill = "MARCA DE AUTO") +
  scale_fill_brewer(palette = "Paired") +  # Usando paleta predefinida
  theme_void() +
  theme(legend.position = "right",
        plot.title = element_text(hjust = 0.5, face = "bold"))
```


Distribucion Marcas de Autos
21.7%
5%
11.7%
20%
21.7%
20%
MARCA DE AUTO
AUDI
RENAULT
BMW
CHEVROLET
FORD
SIN_CONFIRMAR

5 Análisis de la variable “Estatura”.

a. Revisar la variable “ESTATURA” para asegurar que este correctamente importada y limpia. **Rta:**

b. Identificar los datos faltantes o valores inconsistentes. Eliminar esos registros. **Rta:**

c. Generar un poligono de frecuencia y una ojiva para analizar la distribución de la estatura de los clientes y extraer conclusiones. **Rta:**

6 Análisis de la variable "Numero de hijos".

a. Cambiar el nombre de l varaible a "n.hijos" para facilitar su manejo. **Rta:**

b. Identificar problemas con los datos y eliminar valores inconsistentes. **Rta:**

c. Generar una grafica de barras con colores y analizar que se puede concluir sobre el numero de hijos de los clientes. **Rta:**

7 Análisis de la variable “Sexo”.

a. Revisar la consistencia de los datos de la varaible “SEXO” y realizar correcciones en caso de encontrar valores inusuales o inconsistentes. **Rta:**

b. Crea una tabla de frecuencias y extraer conclusiones sobre la proporcion de hombres y mujeres en el conjunto de clientes. **Rta:**

8 Preguntas de investigación".

c. ¿Cuantos clientes con doctorado ganan mas de 2 millones de pesos? **Rta:**

d. ¿Cual es el promedio de salario por cada categoria de variable "MARCA DE AUTO"? **Rta:**