ANALISIS EXPLORATORIO 1 CONCESIONARIO DE AUTOS

Alejandra / 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/t1fe
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?

```
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)

```
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:

```
#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

```r
#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.

```
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 "SIN_CONFIRMAR", se convierte todo mayúsculas.

```
#Tratamiento datos MARCA DE AUTOS
#Eliminar columna 1, 2
datos_base <- datos_base[-c(1,2), ]
#Datos Eliminados
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"
# PERSONA 36 cambio a mayuscula marca
datos_base$`MARCA DE AUTO` <- toupper(datos_base$`MARCA DE AUTO`)
```

b. Crear una 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.

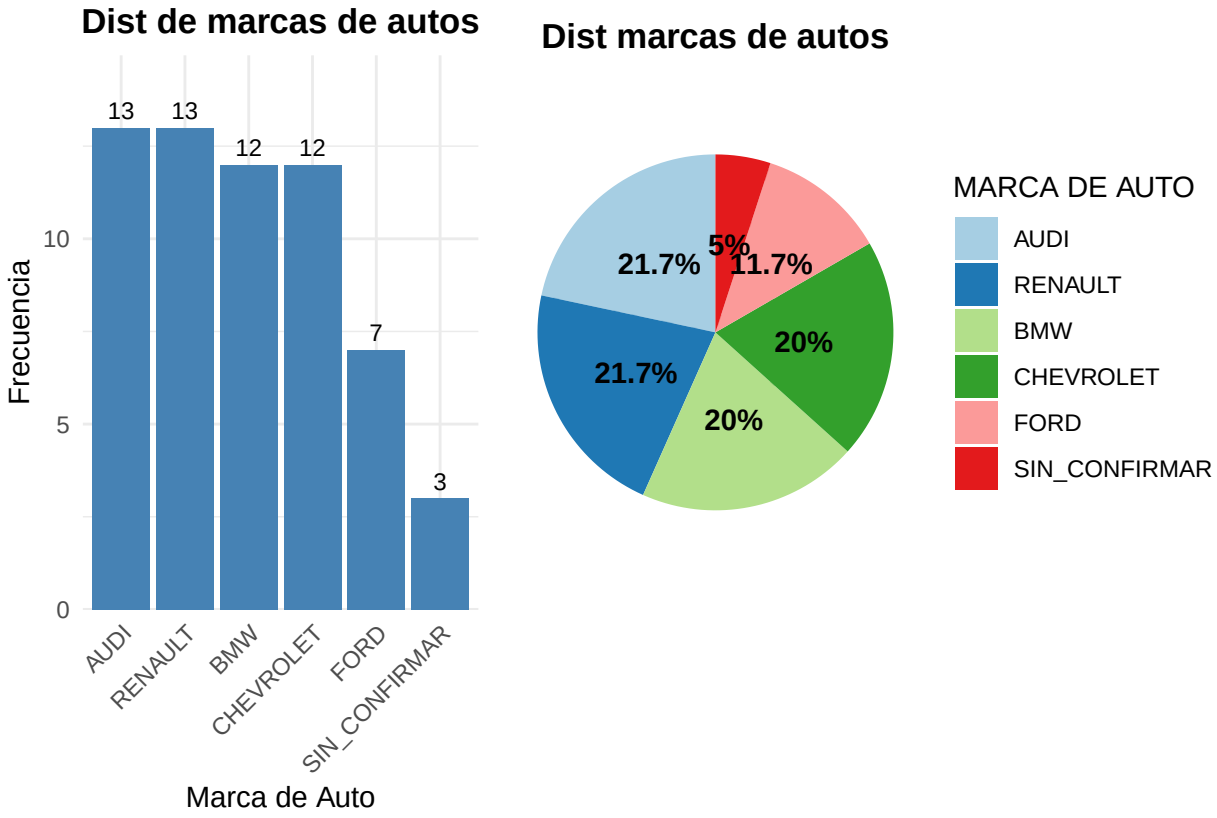
MARCA DE AUTO	Frecuencia
AUDI	13
RENAULT	13
BMW	12
CHEVROLET	12
FORD	7
SIN_CONFIRMAR	3

c. Generar gráficos de barras y de tortas para visualizar la distribución de las marcas de autos y proporcionar una interpretación.

Rta: - Se identifica que las marcas de vehículos preferidas por los clientes del concesionario son **AUDI** y **RENAULT**; se observa una segmentación igual por pares entre las entre AUDI/RENAULT y CHEVROLET/BMW

- 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%.

```
grafico_barras + grafico_pastel   + plot_layout(widths = c(4, 4.5))
```



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%.

4 Análisis de la variable “Edad”.

a. Verificar si la variable “EDAD” está correctamente importada como tipo numérico.

Rta: La variable edad es de tipo **caracter** sin embargo su naturaleza es de tipo Cuantitativa discreta para este conjunto de datos por este motivo se va a relizar la conversión a **numeric** en la sección dedicada a normalización de datos.

```
summary(datos_base$EDAD)
```

```
##    Length     Class      Mode
##        60 character character
```

b. Identificar cualquier valor incorrecto (por ejemplo, texto en lugar de números) y corrija los errores.

Rta: Se encuentran datos incosistententes por formato y **NA** se imputa la media para ambos casos, para el caso de aoutfillers se trabaja con el datos atipico de 10 años para la persona que tenia una edad de 10 años imputando la media.

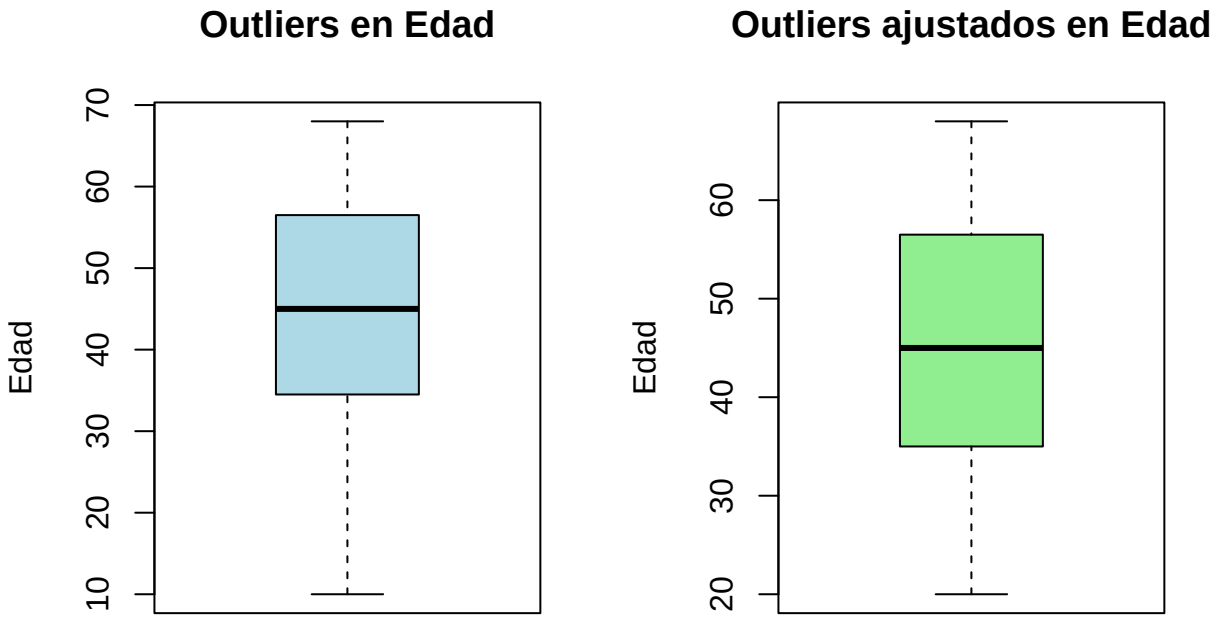
```
#TRATAMIENTO DE DATOS EDAD
#Agrega NA a los datos que no se pueden convertir a numerico
datos_base$EDAD <- as.numeric(datos_base$EDAD)
```

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

```
#Lista las personas que tienen problemas 24 y 28
datos_base[is.na(datos_base$EDAD), c("PERSONA", "EDAD")]
```

```
## # A tibble: 2 x 2
##   PERSONA     EDAD
##   <chr>      <dbl>
## 1 PERSONA 24    NA
## 2 PERSONA 28    NA
```

```
#Se realiza la imputacion de la media a los datos con problema
datos_base$EDAD[is.na(datos_base$EDAD)] <- median(datos_base$EDAD, na.rm = TRUE)
#summary(datos_base$EDAD)
#Datos atipicos
rango_edad <- range(datos_base$EDAD, na.rm = TRUE)
# Mostrar dos gráficos uno al lado del otro
par(mfrow = c(1, 2))  # 1 fila, 2 columnas
boxplot(datos_base$EDAD, main = "Outliers en Edad", ylab = "Edad", col = "lightblue", ylim=rango_edad )
datos_base$EDAD[datos_base$EDAD == 10] <- median(datos_base$EDAD, na.rm = TRUE)
rango_edad <- range(datos_base$EDAD, na.rm = TRUE)
boxplot(datos_base$EDAD, main = "Outliers ajustados en Edad", ylab = "Edad", col = "lightgreen", ylim=ra
```




```
# Restaurar configuración gráfica a 1 gráfico por figura (opcional)
par(mfrow = c(1, 1))
#outliers <- boxplot.stats(datos_base$EDAD)$out ? No funciono
```

c. Realizar un histograma de la variable edad y describir qué patrones se observan entre los clientes (ej. ¿son jóvenes? ¿predomina alguna franja de edad?)

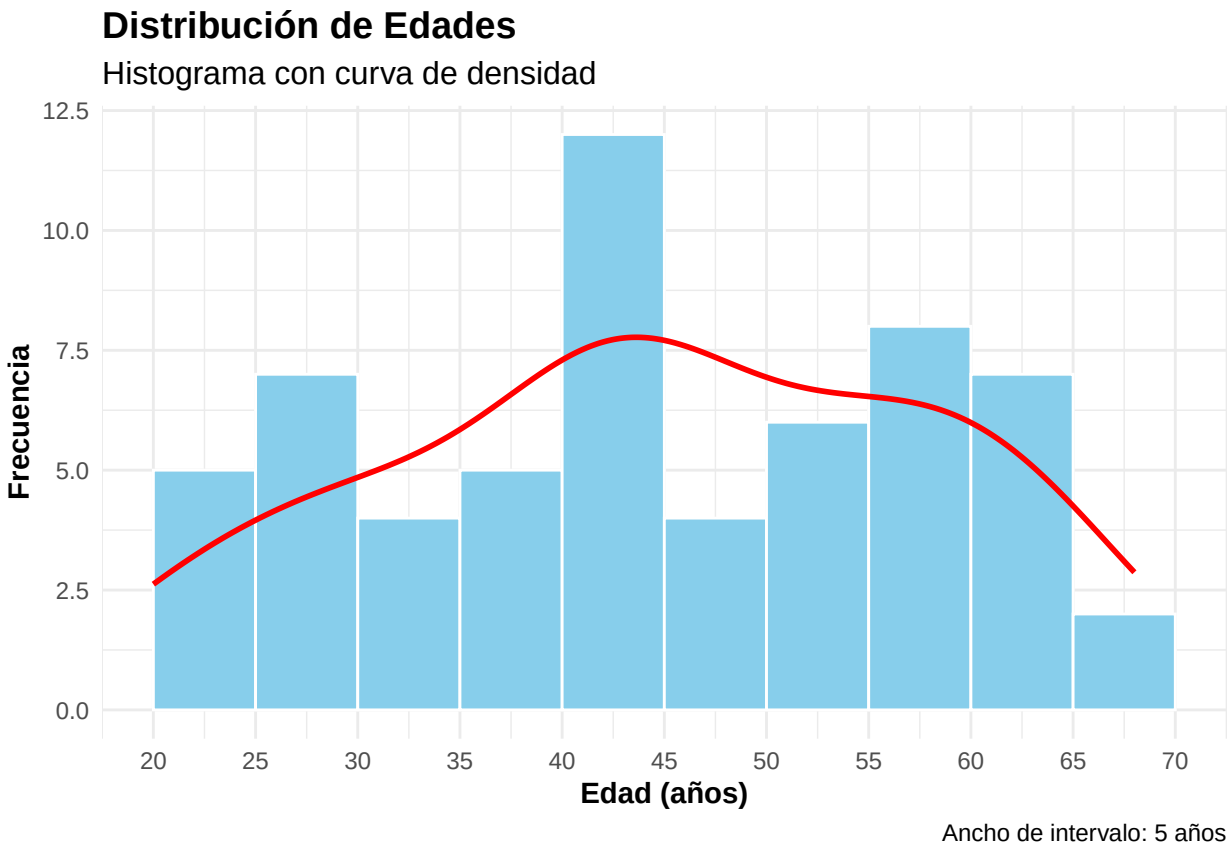
Rta: - La franja de edad en la que se presentan mayor numero de clientes de concesionario estan en la edad de los 40 a los 45 años. - La franja de 55 a 65 años presenta una gran numeros de clientes - El sector joven no presenta una alta concentracion con excepcion de la franja de los 25-30 años

```
# Histograma con curva de densidad para la variable EDAD

ggplot(datos_base, aes(x = EDAD)) +
  geom_histogram(binwidth = 5,
                 fill = "skyblue",
                 color = "white",
                 boundary = 0) +

  # Añadir línea de densidad
  geom_density(aes(y = after_stat(count) * 5),
               color = "red",
               linewidth = 1) +
  labs(title = "Distribución de Edades",
       subtitle = "Histograma con curva de densidad",
```

```
      x = "Edad (años)",
      y = "Frecuencia",
      caption = "Ancho de intervalo: 5 años") +
# Ajustar ejes para mejorar la visualización
scale_x_continuous(breaks = seq(0, max(datos_base$EDAD, na.rm = TRUE) + 5, by = 5)) +
theme_minimal() +
theme(
  plot.title = element_text(face = "bold", size = 14),
  plot.subtitle = element_text(size = 12),
  axis.title = element_text(face = "bold")
)
```



8 Preguntas de investigación.

Utilizando las herramientas de RStudio (filtros, agrupaciones etc), responder a las siguientes preguntas: a. ¿Cuántos clientes tienen mascota?

```
no_clientes_mascota <- datos_base %>%
  filter(!is.na(MASCOTA) & toupper(MASCOTA) == "SI") %>%
  count()
#print(no_clientes_mascota)
```

Rta: El numero de clientes que tiene mascota es: **30**

b. ¿Cuántos clientes mayores de 25 años tienen una maestría?

```
no_clientes_mayores_mascota <- datos_base %>%
  mutate(EDAD = as.numeric(EDAD)) %>%
  filter(!is.na(EDAD) & EDAD > 25) %>%
  filter( `NIVEL ESCOLAR` == "MAESTRÍA") %>%
  count()
    #select(PERSONA, EDAD, `NIVEL ESCOLAR`)
```

Rta: El numero de clientes mayores a 25 años con maestria son **17**

c. ¿Cuántos clientes con doctorado ganan mas de 2 millones de pesos? **Rta:**
d. ¿Cuál es el promedio de salario por cada categoría de la variable “MARCA DE AUTO”?