Estudio de Caso: Análisis del perfil de clientes de un concesionario de autos

Manuel Alejandro Giraldo - Jhonn Jairo Pérez Ramirez

2025-07-23

Tabla de contenido

1 Introducción

En un entorno comercial altamente competitivo, comprender a fondo el perfil de los clientes se convierte en una herramienta clave para optimizar la estrategia comercial de cualquier empresa. Este estudio analiza una base de datos suministrada por un concesionario de autos, con el fin de explorar las características demográficas, educativas y económicas de sus clientes, así como sus preferencias de compra. A través de técnicas de análisis exploratorio de datos en R, se examinan aspectos como edad, nivel educativo, ingresos, marca de vehículo adquirido, tenencia de mascotas e hijos, entre otros. El objetivo es proporcionar una visión clara de los distintos segmentos de clientes y extraer hallazgos que puedan ser aplicados en estrategias de mercadeo, personalización de la oferta y fidelización.

2 Objetivo del análisis

El presente análisis tiene como objetivo caracterizar el perfil de los clientes de un concesionario de vehículos a partir de una base de datos que incluye variables demográficas, socioeconómicas y de comportamiento de compra. Se busca identificar patrones relevantes que permitan comprender mejor el comportamiento de los compradores y ofrecer recomendaciones que puedan orientar decisiones estratégicas en mercadeo, segmentación y oferta de productos.

3 Carga de datos y Exploración inicial de datos

```
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
```

```
library("readxl")
library("ggplot2")
library("paletteer")
library("fdth")
```

```
##
## Adjuntando el paquete: 'fdth'

## The following objects are masked from 'package:stats':
##
##     sd, var
```

Se cargan las librerías **readxl**, **dplyr** y **ggplot2** para poder cargar los datos, analizarlos y graficarlos.

Para realizar el análisis, se utilizó una base de datos en formato Excel (t1fe-tabla_taller.xlsx) que contiene información de los clientes de un concesionario de vehículos. Esta fue importada al entorno de R mediante la función read_excel() del paquete readxl, especificando los valores considerados como faltantes (NA) de forma explícita:

```
data <- read_excel("BASES/t1fe-tabla_taller.xlsx", na = c("", "NA"))
```

```
head(data)
```

```
## # A tibble: 6 x 9
##   PERSONA EDAD  SEXO  ESTATURA `NIVEL ESCOLAR` `MARCA DE AUTO` `NUMERO DE HIJOS`
##   <chr>   <chr> <chr> <chr>    <chr>           <chr>                       <dbl>
## 1 <NA>    <NA>  <NA>  <NA>     <NA>            <NA>                           NA
## 2 <NA>    <NA>  <NA>  <NA>     <NA>            <NA>                           NA
## 3 PERSON~ 21    M     1.54     MAESTRÍA        AUDI                            0
## 4 PERSON~ 26    F     1.55     PROFESIONAL     RENAULT                         5
## 5 PERSON~ 30    F     1.6      DOCTORADO       BMW                             2
## 6 PERSON~ 31    f     1.7      PROFESIONAL     RENAULT                         2
## # i 2 more variables: SALARIO <dbl>, MASCOTA <chr>
```

```
sum(is.na(data))
```

```
## [1] 28
```

Al cargar el archivo, se identificaron dos filas vacías al inicio del documento, lo cual interfiere con la correcta lectura del encabezado y el contenido. Aunque es posible omitir estas filas utilizando el parámetro skip en la función de importación, esta opción eliminaría también la fila correspondiente a los nombres de las columnas. Para evitar este inconveniente, se optó por guardar el encabezado de forma separada y luego unirlo manualmente con el resto de los datos.

Adicionalmente, al explorar la estructura general del dataset, se detectaron 28 valores faltantes distribuidos en distintas variables. Estos fueron tratados como NA mediante un vector de condiciones específicas (na = c("NA")) al momento de la importación para garantizar su correcta codificación.

```
encabezado <- read_excel("BASES/t1fe-tabla_taller.xlsx", n_max=1, col_names=TRUE)

print(names(encabezado))
```

```
## [1] "PERSONA"         "EDAD"            "SEXO"            "ESTATURA"
## [5] "NIVEL ESCOLAR"   "MARCA DE AUTO"   "NUMERO DE HIJOS" "SALARIO"
## [9] "MASCOTA"
```

```
data <- read_excel("BASES/t1fe-tabla_taller.xlsx", skip=2, col_names=FALSE)
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
```

```
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
```

```
head(data)
```

```
## # A tibble: 6 x 9
##   ...1      ...2  ...3  ...4  ...5        ...6    ...7     ...8 ...9
##   <chr>     <chr> <chr> <chr> <chr>       <chr>   <chr>   <dbl> <chr>
## 1 PERSONA 1 21    M     1.54  MAESTRÍA    AUDI    0     1200000 SI
## 2 PERSONA 2 26    F     1.55  PROFESIONAL RENAULT 5     1250000 NO
## 3 PERSONA 3 30    F     1.6   DOCTORADO   BMW     2      900000 NO
## 4 PERSONA 4 31    f     1.7   PROFESIONAL RENAULT 2      800000 NO
## 5 PERSONA 5 35    M     1.71  MAESTRÍA    AUDI    1      950000 NO
## 6 PERSONA 6 65    M     1.8   MAESTRÍA    AUDI    1     2000000 SI
```

```
colnames(data) <- names(encabezado)

head(data)
```

```
## # A tibble: 6 x 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
## 5 PERSON~ 35    M     1.71     MAESTRÍA        AUDI            1
## 6 PERSON~ 65    M     1.8      MAESTRÍA        AUDI            1
## # i 2 more variables: SALARIO <dbl>, MASCOTA <chr>
```

```
tail(data)
```

```
## # A tibble: 6 x 9
##   PERSONA EDAD  SEXO  ESTATURA `NIVEL ESCOLAR` `MARCA DE AUTO` `NUMERO DE HIJOS`
##   <chr>   <chr> <chr> <chr>    <chr>           <chr>           <chr>
## 1 PERSON~ 30    F     1.54     MAESTRÍA        CHEVROLET       2
## 2 PERSON~ 39    M     1.58     MAESTRÍA        AUDI            1
## 3 PERSON~ 34    F     1.6      DOCTORADO       BMW             1
## 4 PERSON~ 24    f     1.7      PROFESIONAL     RENAULT         3
## 5 PERSON~ 20    M     1.71     MAESTRÍA        AUDI            0
## 6 PERSON~ 10    M     1.8      PROFESIONAL     AUDI            0
## # i 2 more variables: SALARIO <dbl>, MASCOTA <chr>
```

```
sum(is.na(data))
```

```
## [1] 6
```

```r
colSums(is.na(data))
```

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

Luego de realizar la limpieza inicial y eliminar las filas vacías al inicio del archivo, se obtuvo un total de 60 registros válidos. En cuanto a los valores faltantes, se identificaron 6 en total, distribuidos de la siguiente manera: uno en la variable SEXO, uno en NIVEL ESCOLAR, dos en MARCA DE AUTO, uno en NÚMERO DE HIJOS y uno en MASCOTA.

```r
str(data)
```

```
## tibble [60 x 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 5
##  $ MASCOTA        : chr [1:60] "SI" "NO" "NO" "NO" ...
```

Al revisar la estructura del dataset, se identificó que la mayoría de las columnas fueron importadas como variables de tipo character (texto), incluyendo algunas que deberían ser numéricas, como ESTATURA, NÚMERO DE HIJOS y EDAD. Por esta razón, es necesario realizar una conversión de tipo de datos, asegurando que estas variables se representaran correctamente como numéricas para permitir su análisis estadístico posterior.

```r
summary(data)
```

```
##    PERSONA              EDAD              SEXO            ESTATURA
##  Length:60          Length:60          Length:60          Length:60
##  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:60          Length:60          Length:60          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
##    MASCOTA
##  Length:60
##  Class :character
##  Mode  :character
```

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

Al utilizar la función `summary()` sobre el dataset, se obtiene un resumen estadístico general. Para las variables de tipo numérico, esta función devuelve valores clave como el mínimo, máximo, media, mediana, así como el primer y tercer cuartil, lo cual facilita una comprensión inicial de la distribución de los datos y la detección de posibles valores atípicos. En cambio, las variables de tipo character solo muestran la cantidad total de registros sin aportar medidas estadísticas.

Dado que al momento de esta revisión las variables EDAD, ESTATURA y NÚMERO DE HIJOS aún no habían sido convertidas a formato numérico, el análisis inicial se centró únicamente en la variable SALARIO. En esta se observó un salario mínimo de $800.000, un máximo de $6.600.000 y una media de aproximadamente $3.286.667. El hecho de que la media sea mayor que la mediana sugiere una asimetría positiva en la distribución, posiblemente generada por la presencia de algunos salarios altos que elevan el promedio por encima del valor central.

3.1 Análisis de la variable "Marca de Auto"

```r
sum(is.na(data$`MARCA DE AUTO`))
```

```
## [1] 2
```

```r
data %>% filter(is.na(data$`MARCA DE AUTO`))
```

```
## # A tibble: 2 x 9
##   PERSONA EDAD  SEXO  ESTATURA `NIVEL ESCOLAR` `MARCA DE AUTO` `NUMERO DE HIJOS`
##   <chr>   <chr> <chr> <chr>    <chr>           <chr>           <chr>
## 1 PERSON~ 68    F     1.65     MAESTRÍA        <NA>            2
## 2 PERSON~ 68    F     1.65     PROFESIONAL     <NA>            3
## # i 2 more variables: SALARIO <dbl>, MASCOTA <chr>
```

Tal como se mencionó anteriormente, la columna MARCA DE AUTO presenta dos valores faltantes, correspondientes a las observaciones de los individuos identificados con los IDs 13 y 49. Al examinar sus registros mediante la función `filter()`, se observó que ambos comparten características similares en variables como edad, nivel educativo, salario y sexo. Las principales diferencias entre ellos radican en el número de hijos y en la presencia o no de mascotas. Esta similitud podría ser útil al momento de decidir cómo imputar el dato faltante, ya que permite buscar patrones en clientes con perfiles comparables.

```r
table(data$`MARCA DE AUTO`, useNA = "ifany")
```

```
##
##      AUDI       BMW       BWM CHEVROLET       FOR      FORD        NA   renault
##        13        11         1        12         1         6         1         1
##   RENAULT      <NA>
##        12         2
```

Al generar la tabla de frecuencias para la variable MARCA DE AUTO, se identificaron tres inconsistencias de escritura que requerían corrección: "BWM" en lugar de "BMW", "FOR" en lugar de "FORD" y "renault" en minúsculas. Para estandarizar estas categorías, se utilizó la función `mutate()` en combinación con `case_when()`, lo cual permitió reemplazar adecuadamente los valores incorrectos.

Además, se detectó un caso en el que el valor “NA” fue importado como texto (string), y no como valor faltante () por R. Junto a este, se encontraron dos valores faltantes reales (). Para garantizar una correcta imputación posterior, fue necesario convertir manualmente el valor “NA” en texto a un NA real, asegurando así la coherencia en el tratamiento de los datos ausentes.

```
#Conversión de "NA" a <NA>
data$`MARCA DE AUTO`[data$`MARCA DE AUTO` %in% c("NA")] <- NA

#Conversión de nombres
data <- data %>%
  mutate(`MARCA DE AUTO` = case_when(
    `MARCA DE AUTO` == "FOR" ~ "FORD",
    `MARCA DE AUTO` == "BWM" ~ "BMW",
    `MARCA DE AUTO` == "renault" ~ "RENAULT",
    TRUE ~ `MARCA DE AUTO`
  ))

#Tabla de frecuencias
table(data$`MARCA DE AUTO`, useNA = "ifany")
```

```
##
##      AUDI       BMW CHEVROLET      FORD   RENAULT      <NA>
##        13        12        12         7        13         3
```

Dado que las tres filas con valores faltantes en la variable MARCA DE AUTO contienen información válida en las demás columnas, no se justifica su eliminación, ya que esto implicaría una pérdida innecesaria de datos potencialmente valiosos. En su lugar, se optó por realizar una imputación por asignación aleatoria entre las categorías más frecuentes. Esta decisión se basa en que, según la tabla de frecuencias, la variable presenta una distribución bimodal, es decir, dos marcas tienen la misma frecuencia máxima. Este enfoque permite preservar la representatividad de las clases más comunes, evitar la introducción de sesgos arbitrarios y, en consecuencia, mejorar la calidad y completitud del conjunto de datos.

```
modas <- c("RENAULT", "AUDI")

data$`MARCA DE AUTO`[is.na(data$`MARCA DE AUTO`)] <-
  sample(modas, sum(is.na(data$`MARCA DE AUTO`)), replace = TRUE)

table(data$`MARCA DE AUTO`, useNA = "ifany")
```

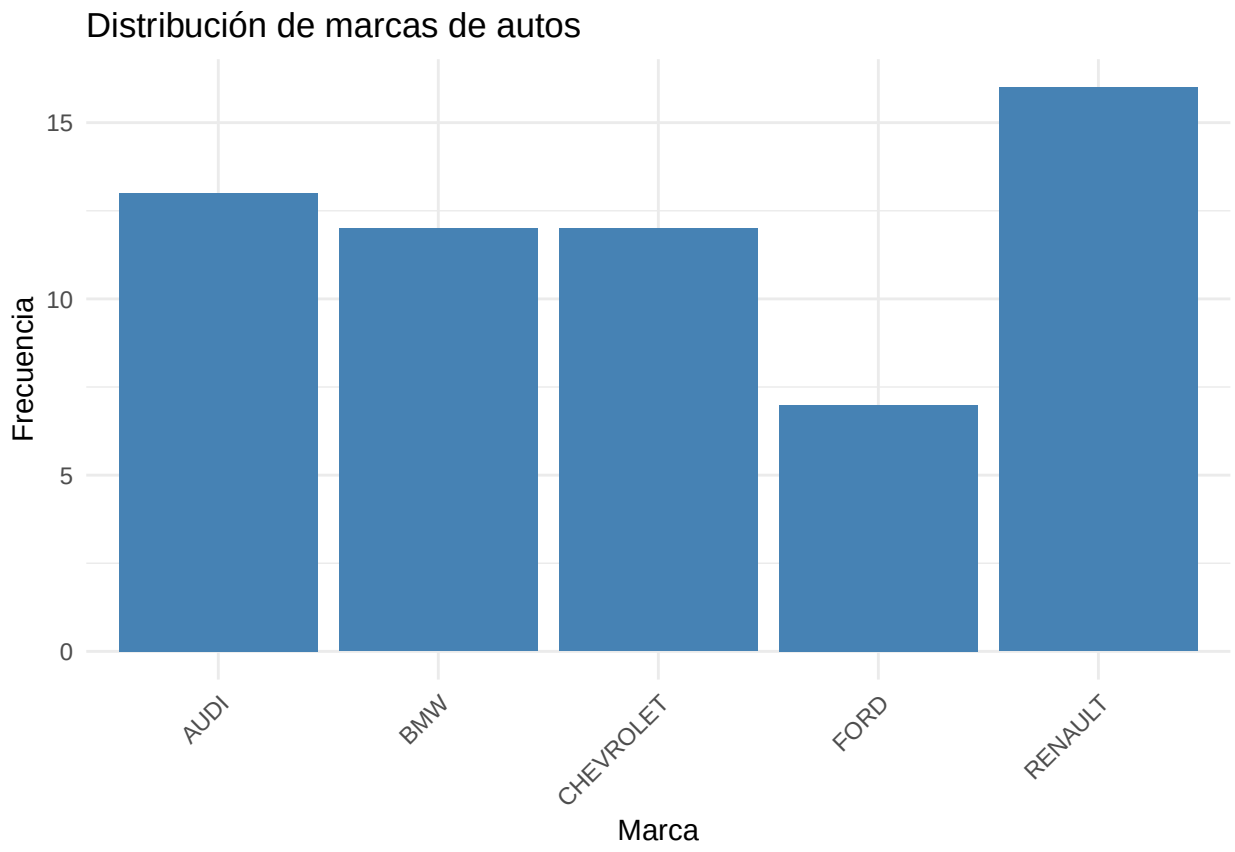
```
##
##      AUDI       BMW CHEVROLET      FORD   RENAULT
##        13        12        12         7        16
```

```
round(prop.table(table(data$`MARCA DE AUTO`, useNA = "ifany")) * 100, 2)
```

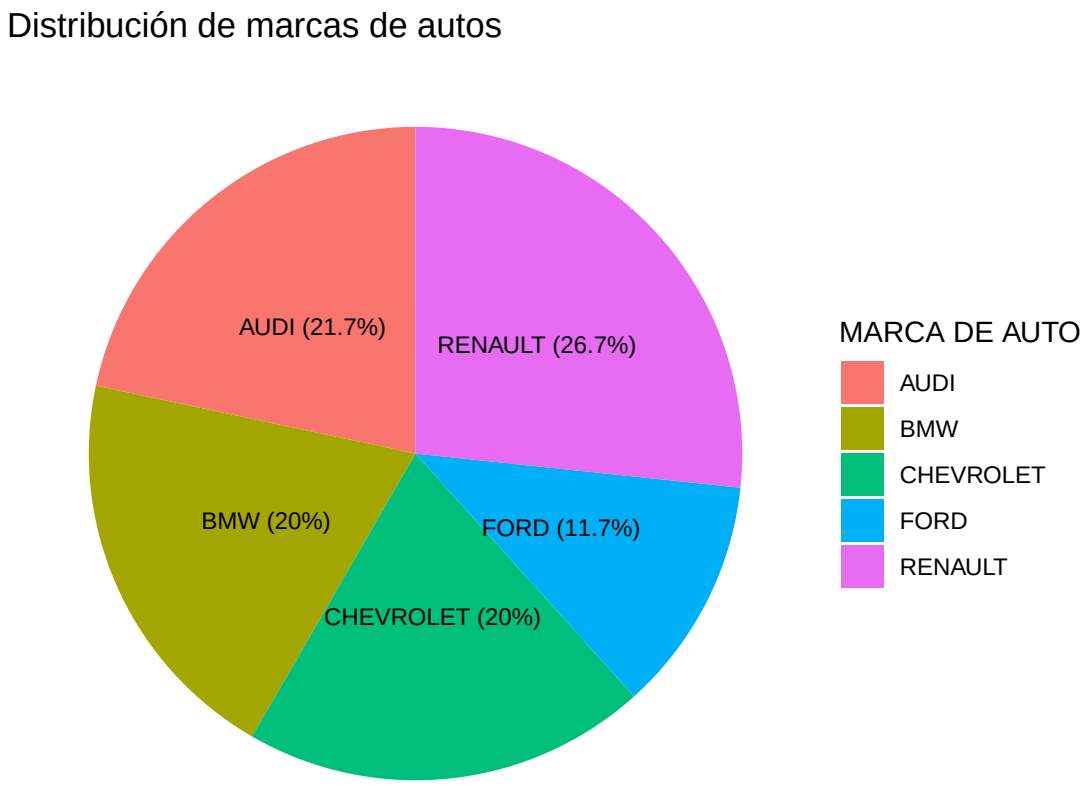
```
##
##      AUDI       BMW CHEVROLET      FORD   RENAULT
##     21.67     20.00     20.00     11.67     26.67
```

Con la imputación realizada, fue posible completar los registros faltantes en la variable MARCA DE AUTO, lo que permitió integrar todos los datos en una única tabla de frecuencias y calcular el porcentaje de participación de cada marca en el total de vehículos vendidos.

```r
ggplot(data=data, aes(x=`MARCA DE AUTO`)) +
  geom_bar(fill = "steelblue") +
  labs(title = "Distribución de marcas de autos", x = "Marca", y = "Frecuencia") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
#Primero calcular la frecuencia relativa
data_marca <- data %>%
  count(`MARCA DE AUTO`) %>%
  mutate(pct = round(100 * n / sum(n), 1),
         label = paste0(`MARCA DE AUTO`, " (", pct, "%)"))

#Grafico de torta
ggplot(data_marca, aes(x = "", y = n, fill = `MARCA DE AUTO`)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start=0) +
  theme_void() +
  labs(title = "Distribución de marcas de autos") +
  geom_text(aes(label = label), position = position_stack(vjust = 0.5), size = 3)
```



A partir del análisis conjunto de la tabla de frecuencias y los gráficos de barras y circular, se concluye que la marca de vehículo más popular entre los clientes es Renault, con un total de 15 unidades vendidas. Le siguen Audi con 14 unidades y tanto BMW como Chevrolet con 12 unidades cada una. En contraste, Ford se posiciona como la marca con la menor preferencia, registrando únicamente 7 unidades vendidas.

Esta distribución sugiere que, dentro del portafolio ofrecido, Renault ha logrado captar un mayor interés. Posiblemente, por su equilibrio entre precio, reputación y características funcionales. Audi y BMW aunque tradicionalmente se asocian a un segmento más premium, también muestran una alta acogida. Por otro lado, la menor preferencia por Ford podría indicar una necesidad de revisar su posicionamiento, precios o presencia en las campañas de promoción actuales.

Estas tendencias pueden servir como base para ajustar estrategias de inventario, precios y marketing, especialmente si se cruzan con otras variables del estudio como edad, nivel educativo y tenencia de mascotas, las cuales pueden influir en la decisión de compra.

3.2 Análisis de la variable "Edad"

Tal como se mencionó durante el proceso de importación, la variable EDAD fue cargada como tipo texto (character), por lo que se requiere su conversión a tipo numérico para permitir su análisis. Esta situación sugiere que podrían existir valores inconsistentes o en formatos no estándar que hayan interferido con la correcta interpretación automática por parte de R.

```
unique(data$EDAD)
```

```
##  [1] "21" "26" "30" "31" "35" "65" "45" "42" "52" "63" "57" "58" "68" "41" "62"
## [16] "53" "51" "40" "60" "56" "NA" "37" "39" "28" "20" "61" "38" "50" "47" "20"
## [31] "46" "34" "24" "10"
```

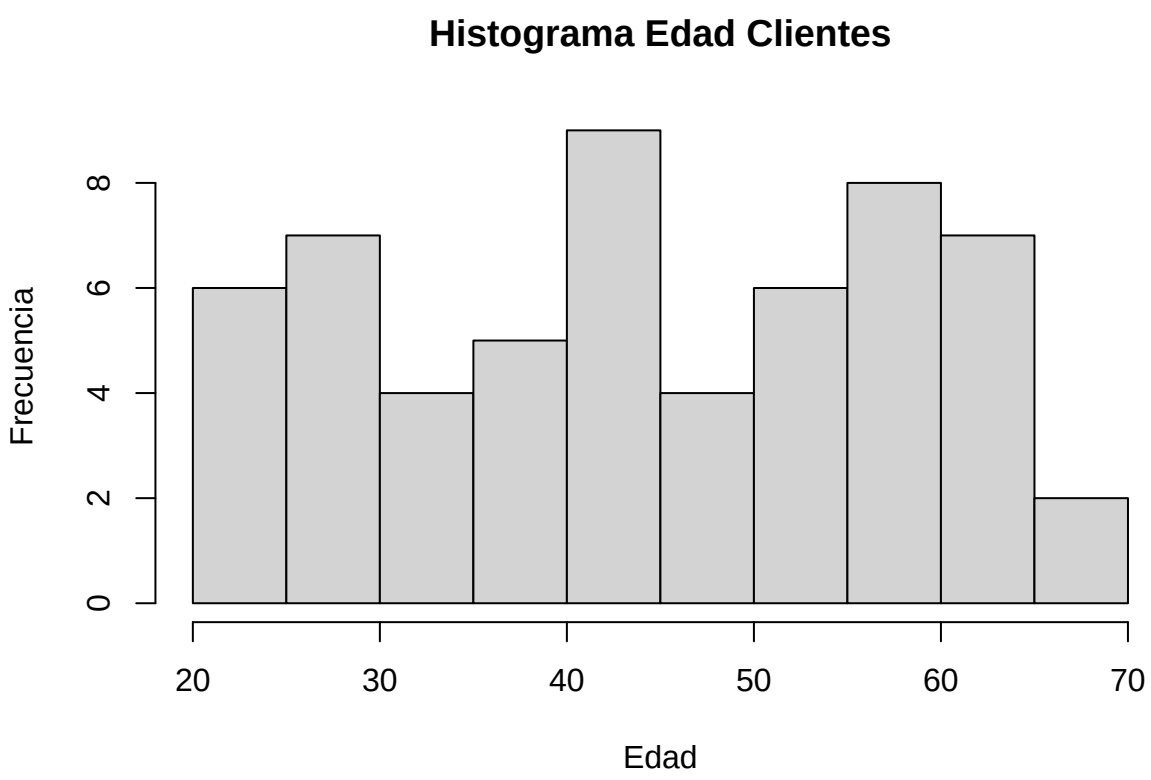
Al examinar los valores únicos de la variable EDAD, se identificaron dos inconsistencias: un "NA" importado como texto (que debe convertirse a un NA real) y un valor "2O" donde la letra O fue digitada en lugar del número 0, por lo que se corrigió a 20. Posteriormente, la variable fue convertida correctamente a tipo numérico.

Además, se detectó un valor atípico de 10 años, el cual no resulta coherente con el contexto del análisis, dado que los menores de edad no pueden adquirir vehículos legalmente. En este caso, se optó por reemplazar dicho valor por la mediana del conjunto, lo cual permite mantener la estabilidad estadística de la variable sin afectar significativamente la distribución general.

```
data$`EDAD`[data$`EDAD` %in% c("NA")] <- NA

data <- data %>%
  mutate(`EDAD` = case_when(
    `EDAD` == "2O" ~ "20",
    `EDAD` == "10" ~ median(data$EDAD),
    TRUE ~ `EDAD`
  ))

data$EDAD <- as.numeric(data$EDAD)

unique(data$EDAD)
```

```
##  [1] 21 26 30 31 35 65 45 42 52 63 57 58 68 41 62 53 51 40 60 56 NA 37 39 28 20
## [26] 61 38 50 47 46 34 24
```

```
hist(data$EDAD, breaks="Sturges",
     main="Histograma Edad Clientes",
     xlab="Edad",
     ylab="Frecuencia")
```



De acuerdo con el histograma, se observa que la mayoría de los compradores se concentran en el rango de edad entre 40 y 60 años, con una mayor densidad en la franja de 50 a 60 años, que representa el grupo más numerosos. También se destaca una frecuencia considerable de clientes entre 20 y 30 años, aunque en menor proporción.

Esta distribución sugiere que los clientes de mediana y avanzada edad constituyen el segmento predominante del concesionario. Este patrón etario podría estar vinculado a una mayor estabilidad laboral y capacidad adquisitiva, así como a una etapa vital en la que la compra de un vehículo se vuelve más estratégica. Ya sea para el transporte familiar, viajes frecuentes o como símbolo de estatus. Por otro lado, la presencia significativa de jóvenes compradores, puede reflejar un nicho emergente, posiblemente motivado por la adquisición de vehículos mediante fianciamiento o por necesidades de movilidad urbana.

3.3 Análisis variable Estatura

```
summary(data$ESTATURA)
```

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

Al revisar la variable ESTATURA, se identificó que fue importada como tipo texto (character), lo cual impide realizar análisis estadísticos directamente. Antes de proceder con su conversión a tipo numérico, es necesario verificar la presencia de valores erróneos o inconsistentes que puedan haber interferido con la correcta interpretación de los datos durante la importación.

```r
unique(data$ESTATURA)
```

```
##  [1] "1.54" "1.55" "1.6"  "1.7"  "1.71" "1.8"  "1.52" "1.51" "1.65" "1.78"
## [11] "1.76" "1.73" "1.81" "1.63" "1.79" "1.59" "1.57" "3.45" "1.50" "1.68"
## [21] "1.53" "1.5"  "1.49" "1.58"
```

Durante la revisión de los valores únicos de la variable ESTATURA, se identificó un valor de 3.45 metros, el cual resulta inconsistente con rangos antropométricos reales. Por lo tanto, se considera un dato atípico y se procede a eliminar la fila correspondiente para evitar distorsiones en el análisis.

Una vez depurado este valor, se realiza la conversión de tipo de dato, transformando la variable de texto (character) a numérico, lo que permite su tratamiento estadístico adecuado.

```r
#Eliminación valor 3.45
data <- data[data$ESTATURA != 3.45, ]

#Conversión a numeric
data$ESTATURA <- as.numeric(data$ESTATURA)

#Resumen de la columna
summary(data$ESTATURA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.490   1.540   1.650   1.655   1.760   1.810
```

Para limpiar la columna ESTATURA, se filtraron todas las filas diferentes al valor atípico de 3.45 metros, eliminando así ese registro inconsistente. Posteriormente, la variable fue convertida de tipo texto a numérico, lo cual permitió volver a generar un resumen estadístico.

En dicho resumen, se observa que la media y la mediana son muy similares, lo que sugiere que la distribución de los datos es ligeramente simétrica. Esta simetría indica una dispersión equilibrada en torno al promedio, sin presencia marcada de sesgos hacia valores altos o bajos.

Para analizar gráficamente esta distribución, se utilizó un polígono de frecuencias generado con ggplot2, añadiendo la capa `geom_freqpoly()` sobre el histograma base. Esta representación permite visualizar la forma general de la distribución de estaturas con mayor claridad y continuidad entre intervalos.

```r
ggplot(data, aes(x=ESTATURA)) + geom_histogram(fill="grey", bins=10) +
  geom_freqpoly(col="red", bins=10) + theme_bw() + labs(
    title = "Distribución de estatura de clientes",
    x = "Estatura (m)",
    y = "Frecuencia"
  )
```

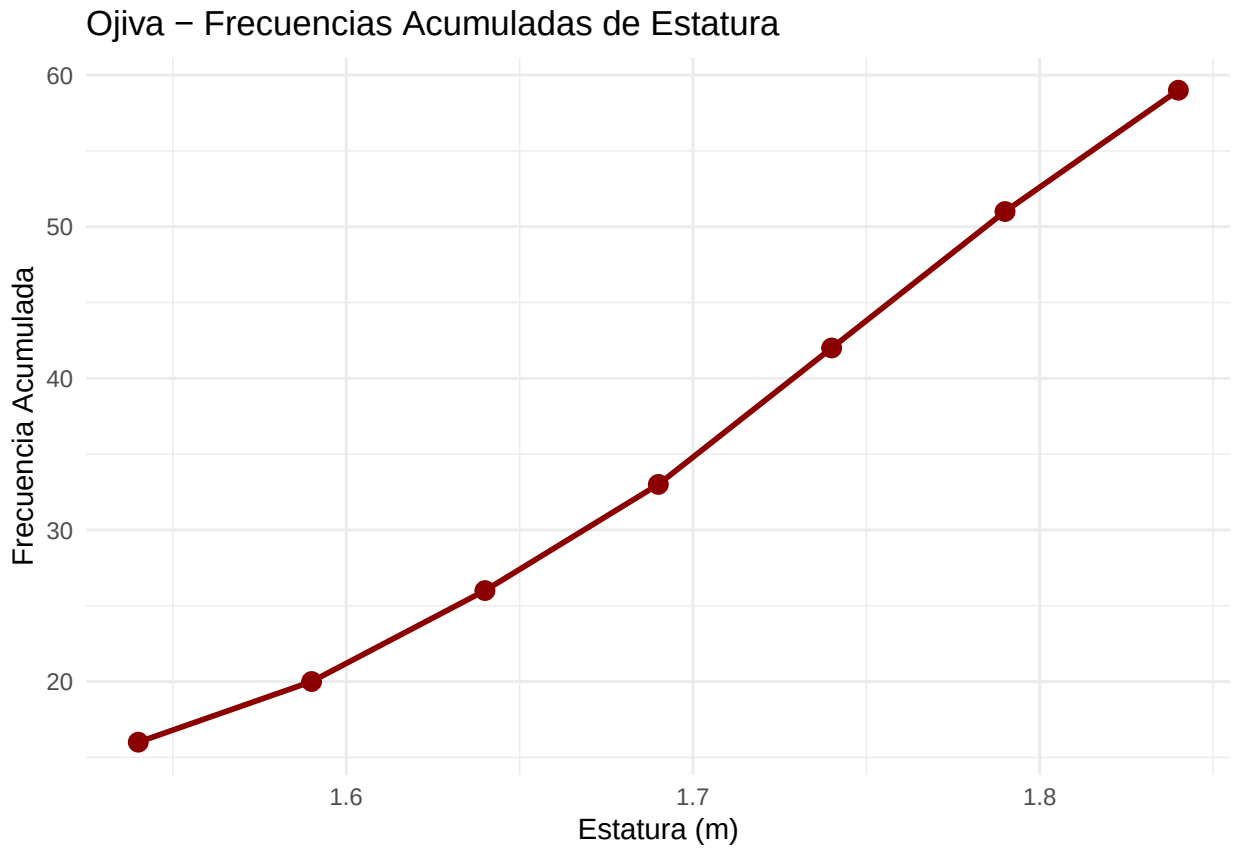
El polígono de frecuencias revela una distribución multimodal, caracterizada por varios picos localizados aproximadamente en 1.50 m y 1.80 m, así como valles entre 1.60 m y 1.70 m. Esta forma no simétrica, que coincide con lo observado en el resumen estadístico de la variable, indica que la distribución no es normal ni unimodal.

Esta variabilidad es coherente con el contexto del conjunto de datos, ya que se trata de una población compuesta por clientes masculinos y femeninos, lo que naturalmente introduce subgrupos diferenciados en términos de estatura.

Adicionalmente, el hecho de que la mayoría de los valores se concentren hacia la parte izquierda del gráfico sugiere una leve asimetría negativa. Cabe señalar que el tamaño de muestra es relativamente pequeño (n=59), lo cual también puede influir en la forma observada de la distribución.

```r
#Ojiva de frecuencias acumuladas

#Creación de los breaks con cortes de 0.05m
breaks <- seq(floor(min(data$ESTATURA, na.rm=TRUE)*100)/100,
  ceiling(max(data$ESTATURA, na.rm=TRUE)*100)/100 + 0.05,
  by = 0.05)

#Crear intervalos y calcular frecuencias acumuladas en un solo paso
hist_data <- hist(data$ESTATURA, breaks=breaks,
  plot=FALSE)

#Crear dataframe para la ojiva
ojiva_data <- data.frame(
  x = hist_data$breaks[-1], #extremos superiores
```

```
  y = c(cumsum(hist_data$counts)) #frecuencias acumuladas
)

#Graficar ojiva
ggplot(ojiva_data, aes(x = x, y = y)) +
  geom_line(color="darkred", linewidth=1) +
  geom_point(color="darkred", size=3) +
  labs(title = "Ojiva - Frecuencias Acumuladas de Estatura",
       x = "Estatura (m)",
       y = "Frecuencia Acumulada") +
  theme_minimal()
```



La ojiva es un gráfico que representa la frecuencia acumulada, mostrando la cantidad de observaciones que se encuentran por debajo o igual a determinados valores. En este caso, se utilizó para analizar la variable ESTATURA.

La curva resultante presenta una forma suave y ascendente, lo que indica que los datos están distribuidos de manera relativamente uniforme a lo largo de los intervalos definidos. La ausencia de saltos abruptos sugiere que no existen acumulaciones anómalas ni concentraciones excesivas en rangos específicos.

Además, se observa que 35 clientes registran estaturas inferiores a 1.70 metros, lo cual proporciona una referencia útil para analizar tendencias generales dentro de la muestra.

3.4 Análisis de la variable "Número de hijos"

Lo primero que se evidenció en esta columna es que su nombre original, NUMERO DE HIJOS, resulta demasiado extenso e incómodo de utilizar durante el análisis y la escritura de código. Por esta razón, se renombró a n.hijos, utilizando funciones del paquete dplyr.

Finalmente, se verificó el cambio listando nuevamente los nombres de las columnas para asegurar que la modificación se haya aplicado correctamente.

```
data <- data %>% rename("n.hijos" = `NUMERO DE HIJOS`)

colnames(data)
```

```
## [1] "PERSONA"       "EDAD"          "SEXO"          "ESTATURA"
## [5] "NIVEL ESCOLAR" "MARCA DE AUTO" "n.hijos"       "SALARIO"
## [9] "MASCOTA"
```

Al igual que otras variables, la columna n.hijos fue importada como tipo texto (character) en lugar de numérico, lo que impide su análisis estadístico. Por esta razón, se hace necesario convertirla al tipo numérico.

Además, se identificó un valor faltante (NA), el cual será tratado mediante imputación con la mediana, una medida robusta frente a valores atípicos.

Antes de realizar dicha imputación, se procedió a explorar los valores únicos de la variable, con el fin de detectar posibles inconsistencias o errores de digitación que puedan distorsionar el análisis posterior.

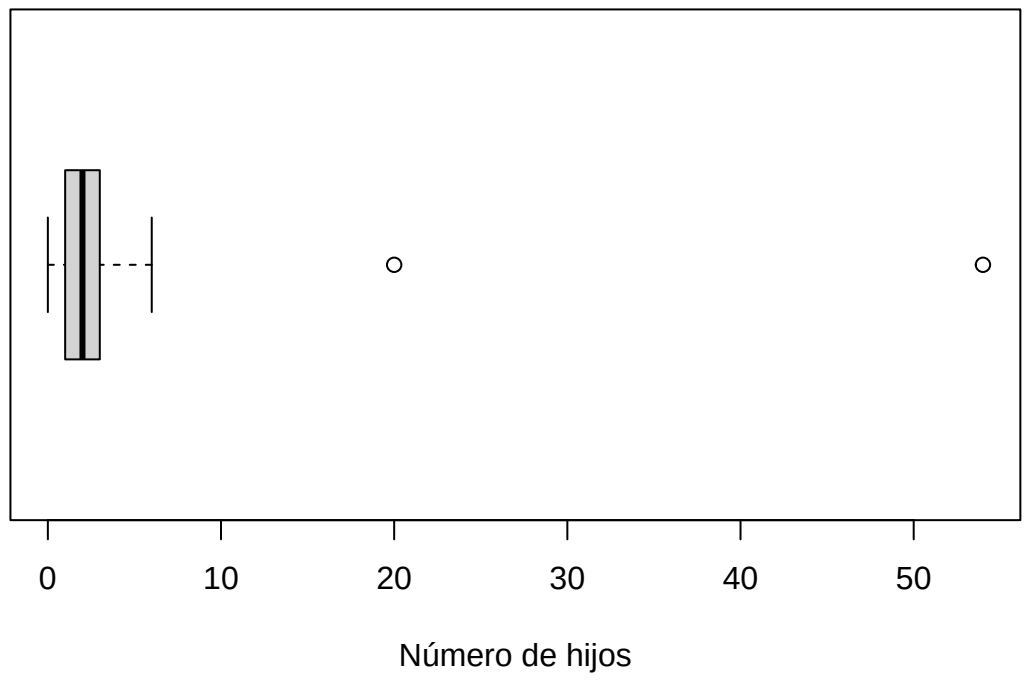
```
#Conversión a numeric
data$n.hijos <- as.numeric(data$n.hijos)
```

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

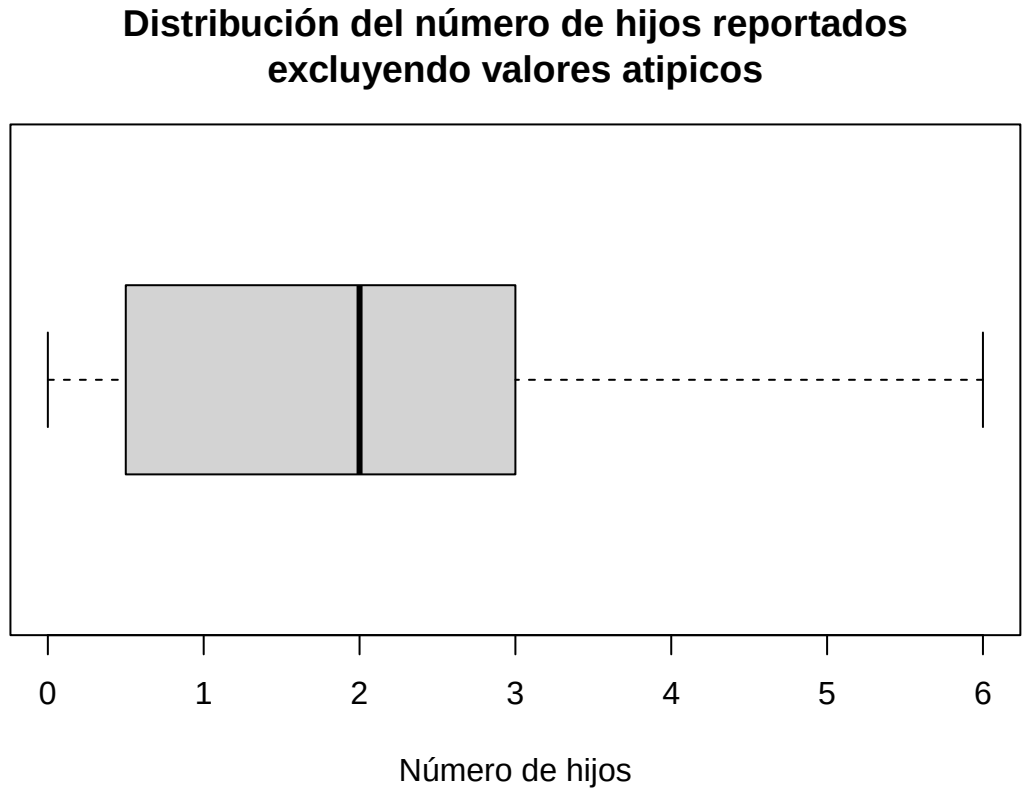
```
#Valores unicos para identificar valores atipicos
unique(data$n.hijos)
```

```
##  [1]  0  5  2  1  4  3 NA 54  6 20
```

```
boxplot(data$n.hijos,
        main="Distribución del número de hijos reportados",
        horizontal = T,
        xlab="Número de hijos")
```




```
nhijos <- data$n.hijos[data$n.hijos < 20]

boxplot(nhijos,
        main="Distribución del número de hijos reportados\nexcluyendo valores atipicos",
        horizontal = T,
        xlab="Número de hijos")
```



Al convertir la columna n.hijos a tipo numérico, R generó un mensaje indicando la introducción de valores NA por coerción, lo cual era esperado debido a que aún no se habían tratado los datos faltantes o inconsistentes.

Al examinar los valores únicos, se identificaron dos observaciones atípicas: 20 y 54. En el caso del 20, es razonable asumir que se trata de un error de digitación, en el que probablemente se ingresó un cero adicional, por lo que se corrigió a 2 hijos.

En cuanto al valor 54, aunque pueden plantearse posibles justificaciones culturales, como vínculos polígamos, matrimonios múltiples o incluso la recolección de datos a nivel familiar (por ejemplo, el total de hijos de un hogar), este valor se encuentra extremadamente alejado del resto de la distribución.

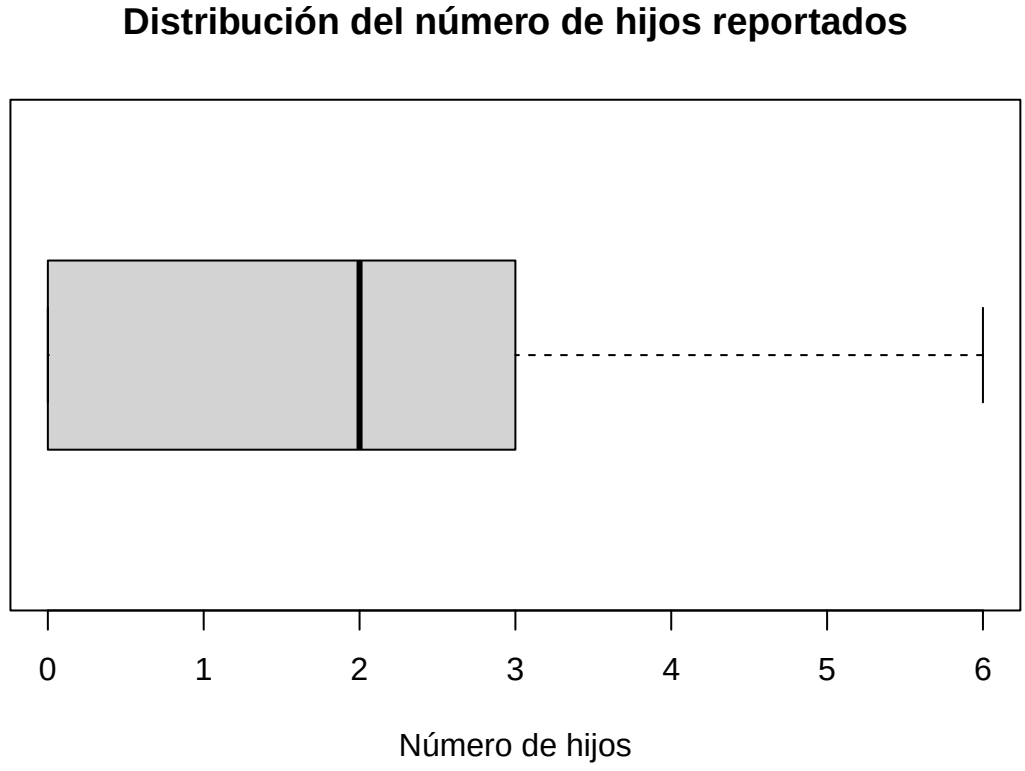
Un análisis mediante boxplot reveló que el número máximo de hijos, excluyendo dicho valor, es 6, el cual ya se ubica por encima del tercer cuartil (Q3), lo que indica una distribución fuertemente sesgada hacia la derecha. En este contexto, el 54 no solo es un dato inusual, sino que distorsiona las medidas de tendencia central, como la media, por lo que se clasifica como un outlier extremo.

Por tanto, se optó por reemplazar este valor por la mediana de la variable, debido a que esta medida es resistente a valores extremos, permitiendo preservar la coherencia estadística del conjunto de datos sin necesidad de eliminar el registro.

Finalmente, los valores faltantes (NA) fueron reemplazados por 0, bajo el supuesto de que su ausencia se debe a clientes sin hijos que no lo reportaron explícitamente.

```
data <- data %>%
  mutate(`n.hijos` = case_when(
    `n.hijos` == 20 ~ 2,
    `n.hijos` == 54 ~ median(data$n.hijos, na.rm=TRUE),
    is.na(`n.hijos`) ~ 0,
    TRUE ~ `n.hijos`
```

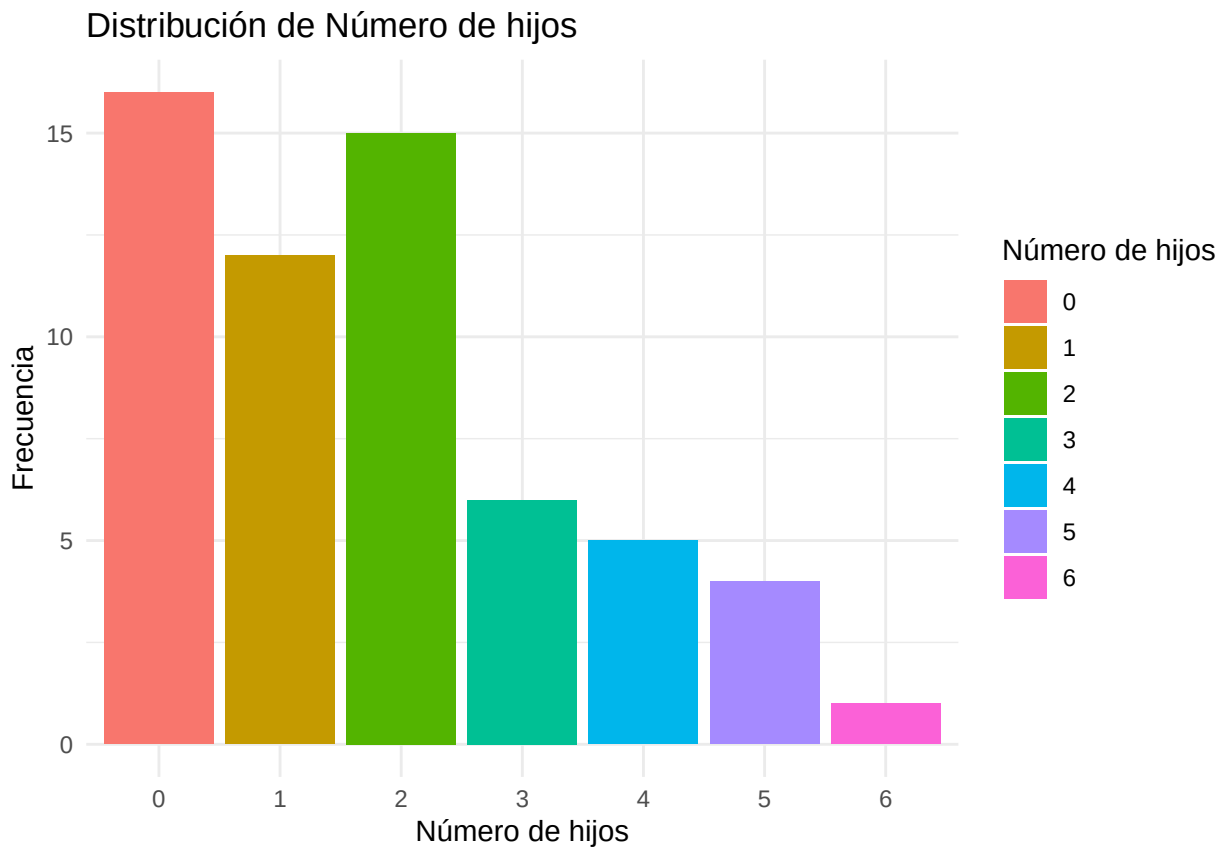
```
  ))

boxplot(data$n.hijos,
        main="Distribución del número de hijos reportados",
        horizontal = T,
        xlab="Número de hijos")
```




```
summary(data$n.hijos)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   2.000   1.797   3.000   6.000
```

```
ggplot(data=data, aes(x= factor(`n.hijos`), fill = factor(`n.hijos`))) +
  geom_bar() +
  labs(title = "Distribución de Número de hijos",
       x = "Número de hijos",
       y = "Frecuencia",
       fill = "Número de hijos") +
  theme_minimal()
```



El gráfico de barras ilustra la distribución del número de hijos entre los clientes del concesionario. Se evidencia que la mayoría no tiene hijos, con una frecuencia cercana a 17 personas, seguida por quienes reportan dos hijos, quienes también representan una proporción considerable de la muestra.

A partir de tres hijos, la frecuencia disminuye de manera progresiva, siendo poco comunes los casos de cinco o más hijos. Esta distribución sugiere que, en esta población, los escenarios más frecuentes son no tener hijos o tener dos, mientras que las familias numerosas son poco representativas.

Este patrón podría estar asociado a factores como la edad promedio de los clientes, su nivel educativo o socioeconómico, o incluso el contexto urbano o rural del cual proviene la muestra. Estas hipótesis podrían ser exploradas con mayor profundidad en análisis posteriores que incorporen variables cruzadas o segmentaciones más específicas.

3.5 Análisis de la variable “Sexo”

La columna SEXO fue importada como carácter (character), lo cual es esperable dado que representa una variable categórica con valores como “H” para hombre y “M” para mujer.

Sin embargo, antes de proceder con cualquier análisis, es fundamental explorar los valores únicos presentes en la columna, con el fin de detectar posibles inconsistencias o variaciones en la codificación (por ejemplo, diferencias en mayúsculas/minúsculas, espacios en blanco, errores tipográficos o categorías adicionales no previstas).

Esta revisión permitirá decidir si es necesario aplicar un proceso de normalización o estandarización, asegurando la coherencia de los datos para futuras segmentaciones o análisis comparativos.

```r
unique(data$SEXO)
```

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

Tras revisar los valores únicos de la columna SEXO, se identificaron variaciones en la notación, incluyendo abreviaciones en inglés como “M” (male) y “F” (female), así como diferencias en el uso de mayúsculas y minúsculas (por ejemplo, “m”, “f”, “HOMBRE”, “mujer”).

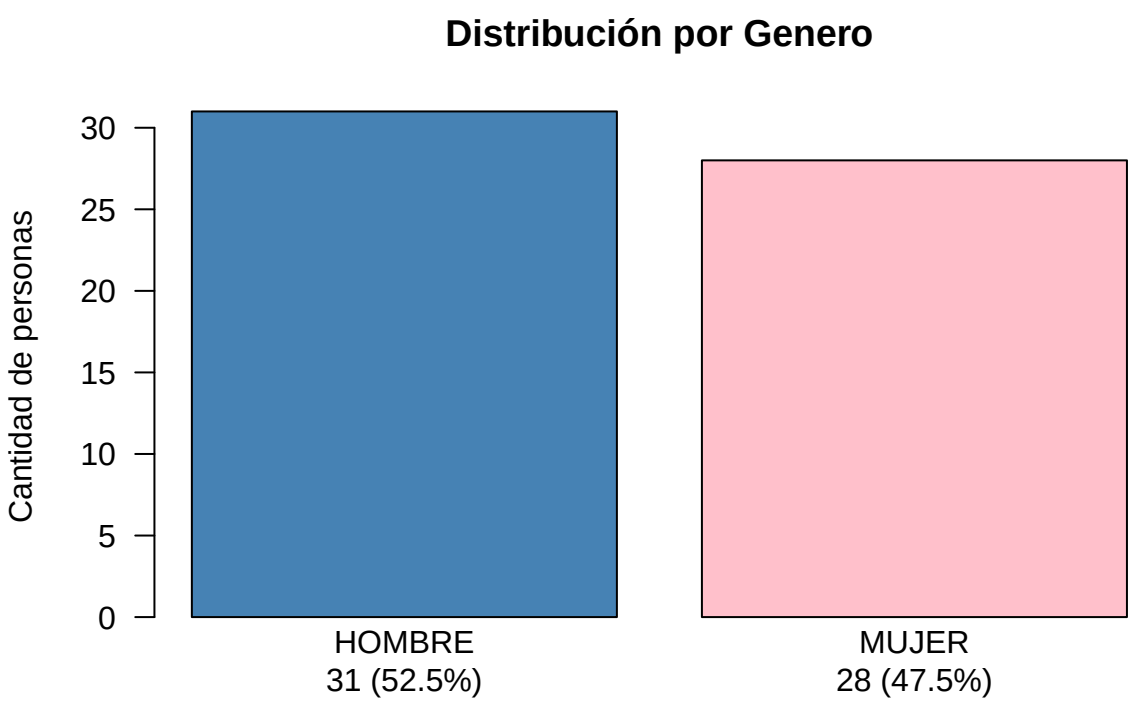
Para garantizar la homogeneidad de los datos y facilitar su análisis posterior, se procedió a realizar una normalización de los valores. Utilizando la función `mutate()` del paquete dplyr, todos los registros fueron recodificados de forma consistente como “HOMBRE” y “MUJER”.

Una vez estandarizados los datos, se generó una tabla de frecuencias para explorar la distribución de género entre los clientes del concesionario y facilitar comparaciones en análisis posteriores.

```r
data <- data %>%
  mutate(
    `SEXO` = case_when(
      `SEXO` == "mujer" ~ "MUJER",
      `SEXO` == "hombre" ~ "HOMBRE",
      `SEXO` == "m" ~ "HOMBRE",
      `SEXO` == "f" ~ "MUJER",
      `SEXO` == "F" ~ "MUJER",
      `SEXO` == "M" ~ "HOMBRE",
      TRUE ~ `SEXO`
    )
  )

frec_sexo = table(data$`SEXO`, useNA = "ifany")
frec_sexo
```

```
##
## HOMBRE  MUJER
##     31     28
```

```r
proporcion <- prop.table(frec_sexo)
porcentaje <- round(proporcion * 100, 1)
etiquetas <- paste0(names(frec_sexo), "\n", frec_sexo, " (", porcentaje, "%)")

barplot(frec_sexo,
        col = c("steelblue", "pink"),
        main = "Distribución por Genero",
        ylab = "Cantidad de personas",
        names.arg = etiquetas,
        las = 1)
```



De esta manera, se identificaron 31 hombres y 28 mujeres entre los 59 registros válidos del dataset, recordando que previamente se eliminó un caso por contener información inconsistente. Esta ligera mayoría masculina podría tener múltiples explicaciones desde una perspectiva económica y sociocultural.

Por un lado, se ha documentado que históricamente los hombres han tenido mayor acceso a empleos formales y a mayores niveles de ingreso, lo cual se traduce en un mayor poder adquisitivo, facilitando la compra de bienes de alto valor como vehículos.

Por otro lado, factores relacionados con los roles de género tradicionales también pueden influir. Culturalmente, el hombre ha sido asociado al papel de proveedor del hogar, y el automóvil puede simbolizar estatus, autonomía y movilidad, reforzando esa narrativa.

No obstante, es importante tener en cuenta que el género del comprador no necesariamente refleja el del usuario principal del vehículo. En muchos casos, los autos pueden registrarse a nombre de un hombre por razones legales, bancarias o sociales, aunque sean utilizados por ambos miembros del hogar. Este tipo de dinámicas podría explorarse con mayor profundidad mediante encuestas directas o variables complementarias.

4 Preguntas de investigación

4.1 ¿Cuántos clientes tienen una mascota?

Para responder esta pregunta, es necesario analizar la columna MASCOTA. Antes de proceder con el análisis descriptivo, se realiza una evaluación inicial de calidad de los datos, con el objetivo de identificar la presencia de valores vacíos, atípicos o inconsistentes.

Esta etapa es fundamental para garantizar que las conclusiones derivadas sean confiables y estén basadas en datos correctamente codificados. En particular, se verificará si existen entradas en blanco, valores como “NA”

o “N/A” interpretados como texto, o categorías adicionales no contempladas originalmente (por ejemplo, diferencias en el uso de mayúsculas o errores de digitación).

```
sum(is.na(data$MASCOTA))
```

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

```
unique(data$MASCOTA)
```

```
## [1] "SI" "NO" NA
```

Al revisar la columna MASCOTA, se identificó un dato vacío, el cual fue tratado como valor faltante (NA). Considerando el contexto y el tipo de variable, se asumió que este dato no fue diligenciado debido a la ausencia de una mascota, por lo que se decidió imputarlo con la categoría “NO”.

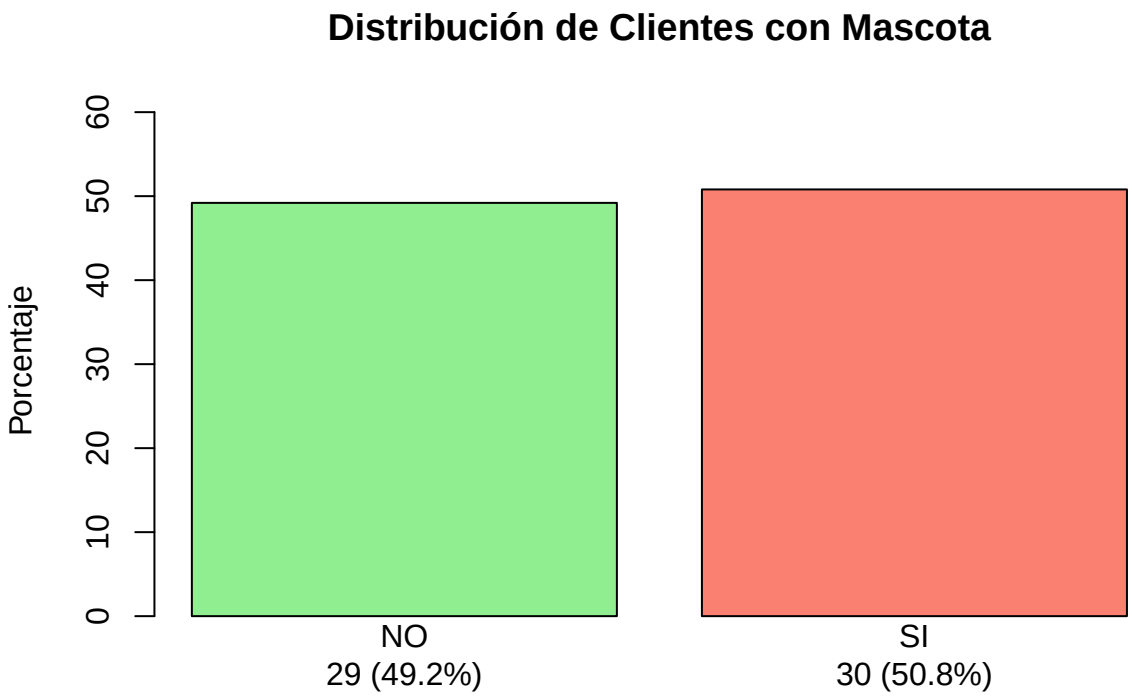
Esta decisión busca preservar la completitud del dataset sin introducir sesgos significativos, especialmente teniendo en cuenta que no existe evidencia que sugiera un error sistemático en la recolección de este dato.

Una vez corregido este valor, se procedió a calcular la frecuencia absoluta de clientes con y sin mascota, lo cual permite identificar tendencias generales de tenencia de mascotas entre los compradores del concesionario.

```
data <- data %>%
  mutate(`MASCOTA` = case_when(
    is.na(`MASCOTA`) ~ "NO",
    TRUE ~ `MASCOTA`
  ))

table(data$`MASCOTA`, useNA = "ifany")
```

```
##
## NO SI
## 29 30
```

```
frec_mascota = table(data$`MASCOTA`, useNA="ifany")
prop_mascota = round(prop.table(frec_mascota) * 100, 1)
etiquetas <- paste0(names(frec_mascota), "\n", frec_mascota, " (", prop_mascota, "%)")

barplot(prop_mascota,
        main = "Distribución de Clientes con Mascota",
        ylab = "Porcentaje",
        col = c("lightgreen", "salmon"),
        ylim = c(0, max(prop_mascota) + 10),
        names.arg = etiquetas)
```



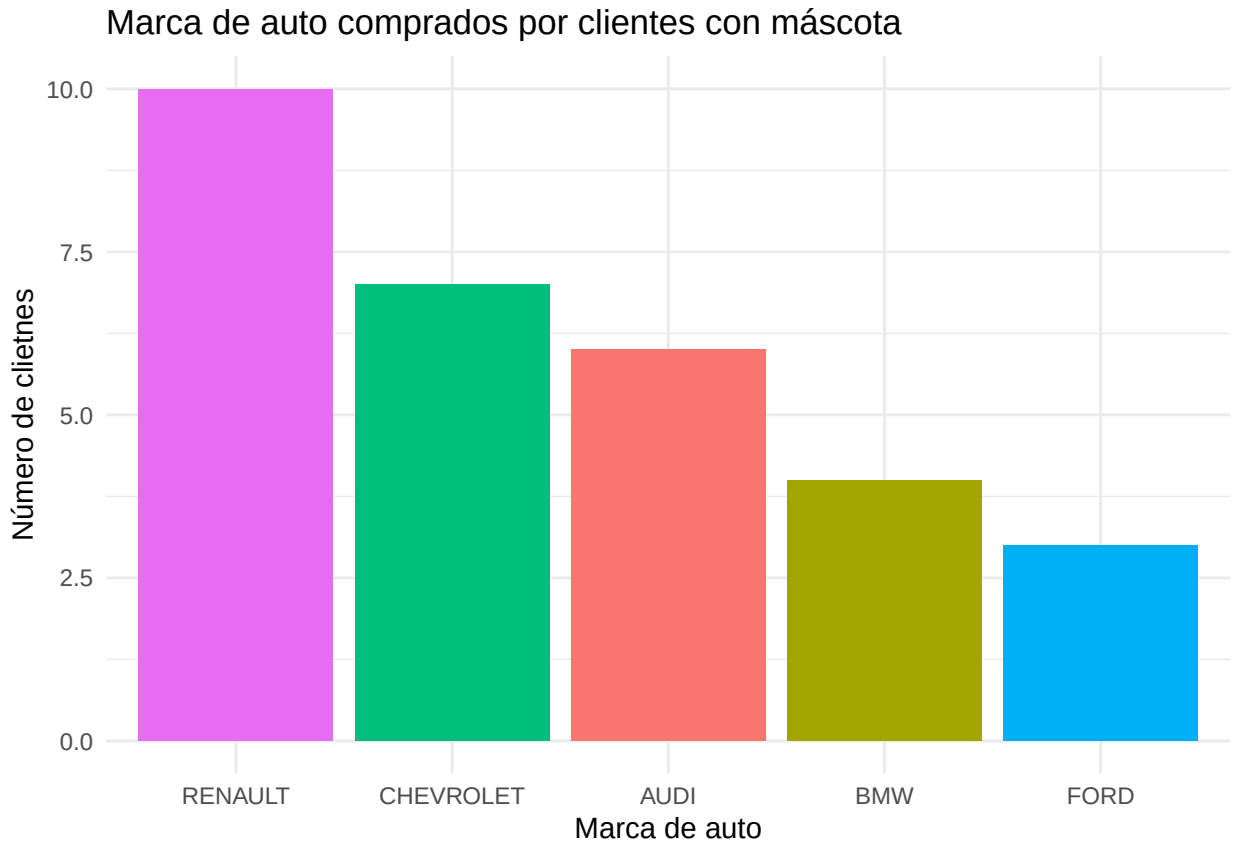
De esta manera, se identificó que 30 clientes del concesionario tienen mascotas, lo cual representa aproximadamente la mitad de la muestra. Este dato es relevante, ya que la tenencia de mascotas puede influir directamente en las decisiones de compra de vehículos.

Por ejemplo, los clientes con mascotas tienden a valorar características específicas, como un mayor espacio interior, facilidad de limpieza, baúles amplios, ventanas traseras seguras o incluso accesorios especiales para el transporte de animales.

Esta información resulta útil para segmentar el mercado y personalizar estrategias comerciales, como campañas promocionales dirigidas, recomendaciones de modelos más adecuados para este perfil o servicios complementarios (accesorios, mantenimientos específicos, seguros para mascotas, entre otros). Aprovechar este tipo de datos permite optimizar la oferta según las necesidades reales de los clientes, fortaleciendo la relación con este segmento y potenciando las ventas.

```
#Filtrar solo quienes tienen mascota
clientes_con_mascota <- data %>%
  filter(`MASCOTA` == "SI")

#Contar tipos de vehiculos entre quienes tienen mascota
conteo_vehiculos <- clientes_con_mascota %>%
  group_by(`MARCA DE AUTO`) %>%
  summarise(Frecuencia = n()) %>%
  arrange(desc(Frecuencia))

#Gráfica
ggplot(conteo_vehiculos, aes(x=reorder(`MARCA DE AUTO`, -Frecuencia),
                             y=Frecuencia, fill = `MARCA DE AUTO`)) +
```

```
  geom_bar(stat = "identity") +
  labs(title = "Marca de auto comprados por clientes con máscota",
       x = "Marca de auto", y = "Número de clietnes") +
  theme_minimal() +
  theme(legend.position = "none")
```



Para comprender las preferencias de los clientes con mascotas en relación con la marca de vehículo adquirido, se realizó un filtrado de la base de datos seleccionando únicamente aquellos registros cuya variable MASCOTA tiene el valor "SI". Este subconjunto permitió identificar de manera específica el comportamiento de este grupo poblacional.

Posteriormente, se agruparon los datos por la variable MARCA DE AUTO y se contó el número de clientes en cada categoría, ordenando el resultado de mayor a menor frecuencia. Esta operación, se realizó con el fin de identificar las marcas más elegidas entre los clientes con mascotas.

Finalmente, se elaboró un gráfico de barras en el que se visualiza esta distribución, con cada barra representando el número de clientes por marca. Se utilizó la función ggplot de la librería ggplot2, incluyendo colores diferenciados para cada marca, aunque se ocultó la leyenda para evitar redundancia visual.

Al analizar las marcas de vehículos adquiridas por los clientes que tienen mascota, se observa que Renault lidera con 10 unidades vendidas, seguida por Chevrolet (7) y Audi (6).

Esta distribución sugiere una posible preferencia hacia marcas que ofrecen modelos con mayor espacio interior, practicidad o funcionalidades compatibles con la movilidad de animales de compañía.

A un nivel más estratégico, estos resultados permiten inferir que los clientes con mascotas podrían inclinarse por vehículos familiares o utilitarios, lo cual representa una oportunidad para segmentar campañas de marketing, personalizar recomendaciones comerciales o destacar características de modelos que respondan mejor a este perfil.

Aprovechar esta información permite al concesionario alinear su oferta con las necesidades reales del cliente, mejorando la eficacia de sus acciones comerciales y fortaleciendo la fidelización de este segmento.

4.2 ¿Cuántos clientes mayores de 25 años tienen una maestría?

Para poder responder esta pregunta, es necesario explorar y normalizar la columna NIVEL ESCOLAR, ya que el análisis de cualquier variable requiere trabajar con datos consistentes y estandarizados.

```
#Valores unicos
unique(data$`NIVEL ESCOLAR`)
```

```
## [1] "MAESTRÍA"    "PROFESIONAL" "DOCTORADO"   "PhD"         NA
```

```
#Número de datos vacíos
sum(is.na(data$`NIVEL ESCOLAR`))
```

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

```
data <- data %>%
  mutate(`NIVEL ESCOLAR` = case_when(
    is.na(`NIVEL ESCOLAR`) ~ "NINGUNO",
    `NIVEL ESCOLAR` == "PhD" ~ "DOCTORADO",
    TRUE ~ `NIVEL ESCOLAR`
  ))

clientes_maestria <- data %>%
  filter(`EDAD` >= 25 & `NIVEL ESCOLAR` == "MAESTRÍA")

nrow(clientes_maestria)
```

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

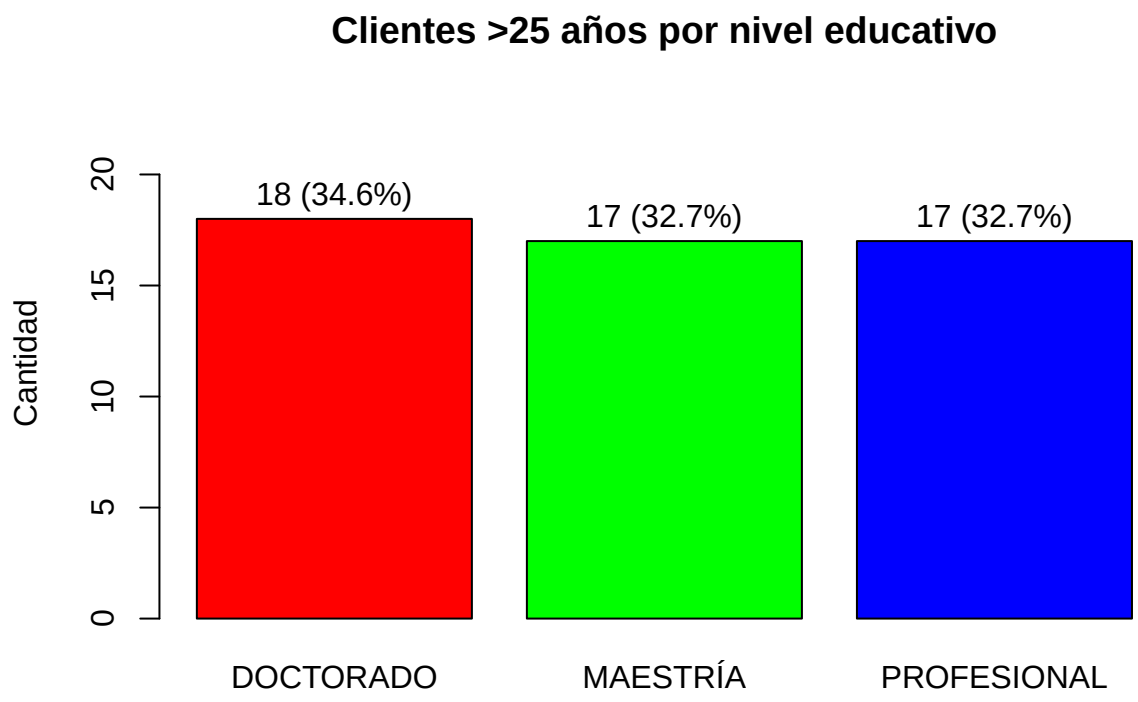
Al revisar los valores únicos, se observa que existen inconsistencias en la forma de registrar los niveles educativos, como diferencias de formato (uso de mayúsculas, abreviaciones como PhD, presencia de valores vacíos o errores de tipeo).

Por esta razón, se procedió a realizar una normalización utilizando la función `mutate()` junto con `case_when()`, estandarizando categorías como "PhD" a "DOCTORADO", tratando los valores faltantes como "NINGUNO", y homogenizando el resto de categorías. Este paso es fundamental para asegurar la coherencia del análisis estadístico y evitar errores de interpretación en pasos posteriores.

```
#Conteo y porcentaje de cada nivel escolar
data %>%
  filter(`EDAD` >= 25) %>%
  count(`NIVEL ESCOLAR`) %>%
  mutate(porcentaje = round(n / sum(n) * 100, 2))
```

```
## # A tibble: 3 x 3
##   `NIVEL ESCOLAR`     n porcentaje
##   <chr>           <int>      <dbl>
## 1 DOCTORADO          18       34.6
## 2 MAESTRÍA           17       32.7
## 3 PROFESIONAL        17       32.7
```

```r
edu_25 = data %>%
  filter(`EDAD` > 25)

freq_educacion = table(edu_25$`NIVEL ESCOLAR`)
porcentaje <- round(prop.table(freq_educacion) * 100, 1)
etiquetas <- paste0(freq_educacion, " (", porcentaje, "%)")

bp <- barplot(freq_educacion,
  main = "Clientes >25 años por nivel educativo",
  ylab = "Cantidad",
  col = rainbow(length(freq_educacion)),
  ylim = c(0, max(freq_educacion) + 5))
text(x = bp, y = freq_educacion + 1, labels = etiquetas)
```




```r
#Distribución por género dentro del grupo con maestría
data %>%
  filter(`EDAD` >= 25 & `NIVEL ESCOLAR` == "MAESTRÍA") %>%
  count(SEXO)
```

```
## # A tibble: 2 x 2
##   SEXO       n
##   <chr>  <int>
## 1 HOMBRE     8
## 2 MUJER      9
```

```r
#Marcas de auto preferidas por este grupo
data %>%
  filter(EDAD >= 25 & `NIVEL ESCOLAR` == "MAESTRÍA") %>%
  count(`MARCA DE AUTO`) %>%
  arrange(desc(n))
```

```
## # A tibble: 5 x 2
##   `MARCA DE AUTO`     n
##   <chr>           <int>
## 1 AUDI                7
## 2 BMW                 4
## 3 CHEVROLET           2
## 4 FORD                2
## 5 RENAULT             2
```

```r
#Salario promedio
data %>%
  filter(EDAD >= 25, `NIVEL ESCOLAR` == "MAESTRÍA") %>%
  summarise(salario_promedio = mean(SALARIO, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   salario_promedio
##              <dbl>
## 1         2976471.
```

```r
#Datos demográficos (hijos, mascotas y estatura)
data %>%
  filter(EDAD >= 25 & `NIVEL ESCOLAR` == "MAESTRÍA") %>%
  summarise(
    promedio_hijos = round(mean(n.hijos, na.rm = TRUE), 0),
    porcentaje_con_mascota = round(mean(MASCOTA == "SI") * 100, 2),
    estatura_promedio = round(mean(ESTATURA, na.rm = TRUE), 2)
  )
```

```
## # A tibble: 1 x 3
##   promedio_hijos porcentaje_con_mascota estatura_promedio
##            <dbl>                  <dbl>             <dbl>
## 1              2                   41.2               1.7
```

Se identificaron 17 clientes mayores de 25 años con nivel educativo de maestría, lo que representa el 32,69% del total de clientes del concesionario. De este grupo, 9 son mujeres y 8 hombres, lo que indica una distribución relativamente equitativa por género. Adicionalmente, 17 clientes son profesionales y 18 tienen doctorado.

En cuanto a sus preferencias vehiculares, la marca más elegida por este segmento es Audi, seguida por BMW y Chevrolet, lo cual podría reflejar una mayor valoración por marcas asociadas con prestigio, calidad y tecnología. Esto se alinea con el hecho de que presentan un salario promedio de aproximadamente 3 millones, lo cual les otorga un mayor poder adquisitivo frente a otros segmentos.

Otros datos de interés incluyen que este grupo tiene un promedio de 2 hijos por persona, una estatura media de 1.70 metros, y que el 41% de ellos convive con una mascota. Estos elementos podrían ser útiles para segmentar ofertas familiares o servicios complementarios como accesorios para mascotas o vehículos con mayor espacio interior.

Este grupo representa un perfil valioso para el concesionario, al conjugar alto nivel educativo, ingresos medio-altos y una preferencia por marcas de gama media-alta. Orientar campañas publicitarias o beneficios especiales hacia este tipo de clientes podría traducirse en una mayor fidelización y mayores ingresos por ventas de vehículos premium.

4.3 ¿Cuántos clientes con doctorado ganan más de 2 millones de pesos?

```
clientes_doctorado <- data %>%
  filter(`NIVEL ESCOLAR` == "DOCTORADO" & SALARIO > 2000000)

nrow(clientes_doctorado)
```

```
## [1] 15
```

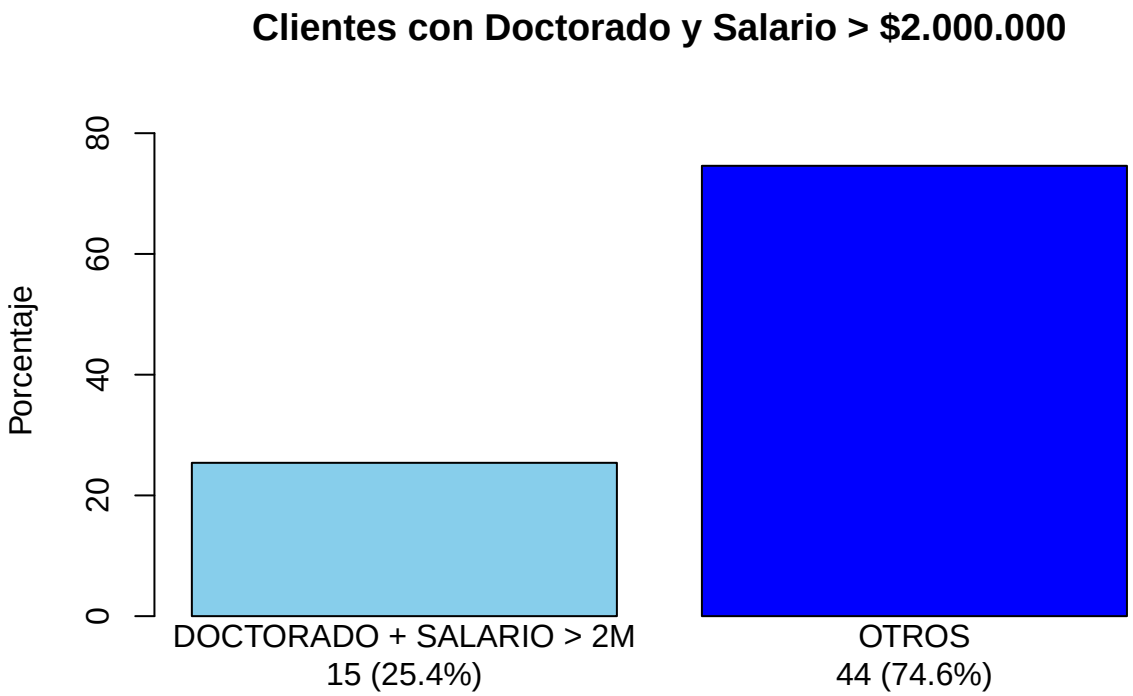
```
data_doc <- toupper(trimws(data$`NIVEL ESCOLAR`))
data_doc$DOCTORADO_ALTO_SALARIO <-  ifelse(
  data$`NIVEL ESCOLAR` == "DOCTORADO" &
    data$SALARIO > 2000000, "DOCTORADO + SALARIO > 2M", "OTROS")
```

```
## Warning in data_doc$DOCTORADO_ALTO_SALARIO <- ifelse(data$`NIVEL ESCOLAR` == :
## Realizando coercion de LHD a una lista
```

```
tabla_doctorado <- table(data_doc$DOCTORADO_ALTO_SALARIO)
porcentaje_doctorado <- round(prop.table(tabla_doctorado) * 100, 1)
etiquetas <- paste0(
  names(tabla_doctorado),
  "\n", tabla_doctorado,
  " (", porcentaje_doctorado, "%)")
print(tabla_doctorado)
```

```
##
## DOCTORADO + SALARIO > 2M                   OTROS
##                       15                      44
```

```
barplot(porcentaje_doctorado,
        main = "Clientes con Doctorado y Salario > $2.000.000",
        ylab = "Porcentaje",
        col = c("skyblue", "blue"),
        names.arg = etiquetas,
        ylim = c(0, max(porcentaje_doctorado) + 10))
```




```
data %>%
  filter(`NIVEL ESCOLAR` == "DOCTORADO" & SALARIO > 2000000) %>%
  count(SEXO)
```

```
## # A tibble: 2 x 2
##   SEXO       n
##   <chr>  <int>
## 1 HOMBRE     9
## 2 MUJER      6
```

```
data %>%
  filter(`NIVEL ESCOLAR` == "DOCTORADO" & SALARIO > 2000000) %>%
  count(`MARCA DE AUTO`) %>%
  arrange(desc(n))
```

```
## # A tibble: 5 x 2
##   `MARCA DE AUTO`     n
##   <chr>           <int>
## 1 RENAULT             6
## 2 BMW                 4
## 3 AUDI                2
## 4 CHEVROLET           2
## 5 FORD                1
```

```r
#Datos demográficos (hijos, mascotas y estatura)
data %>%
  filter(`NIVEL ESCOLAR` == "DOCTORADO" & SALARIO > 2000000) %>%
  summarise(
    edad_promedio = round(mean(EDAD, na.rm=TRUE), 0),
    promedio_hijos = round(mean(n.hijos, na.rm = TRUE), 0),
    porcentaje_con_mascota = round(mean(MASCOTA == "SI") * 100, 2),
    estatura_promedio = round(mean(ESTATURA, na.rm = TRUE), 2)
  )
```

```
## # A tibble: 1 x 4
##   edad_promedio promedio_hijos porcentaje_con_mascota estatura_promedio
##           <dbl>          <dbl>                  <dbl>             <dbl>
## 1            46              1                   66.7              1.63
```

Se identificaron 15 clientes con nivel educativo de doctorado que perciben un salario superior a dos millones de pesos mensuales. Esta población se compone de 9 hombres y 6 mujeres, con una edad promedio de 46 años.

En cuanto a preferencias vehiculares, se evidencia una fuerte inclinación hacia marcas reconocidas. Renault lidera con 6 unidades vendidas, seguida por BMW (4), Audi (2), Chevrolet (2) y Ford (1).

El perfil promedio de este grupo incluye una estatura de 1.63 metros, un promedio de un hijo por persona y un índice de tenencia de mascotas del 66.67%, lo que refuerza la idea de un estilo de vida familiar, profesional y con cierto nivel de estabilidad.

Estos datos sugieren que los clientes con doctorado constituyen un segmento con alto poder adquisitivo y preferencias por vehículos de gama media-alta. Este perfil puede ser relevante para diseñar campañas comerciales enfocadas en clientes con alto nivel académico, destacando modelos con tecnología avanzada, confort, seguridad o prestigio de marca, alineados con sus expectativas y estilo de vida.

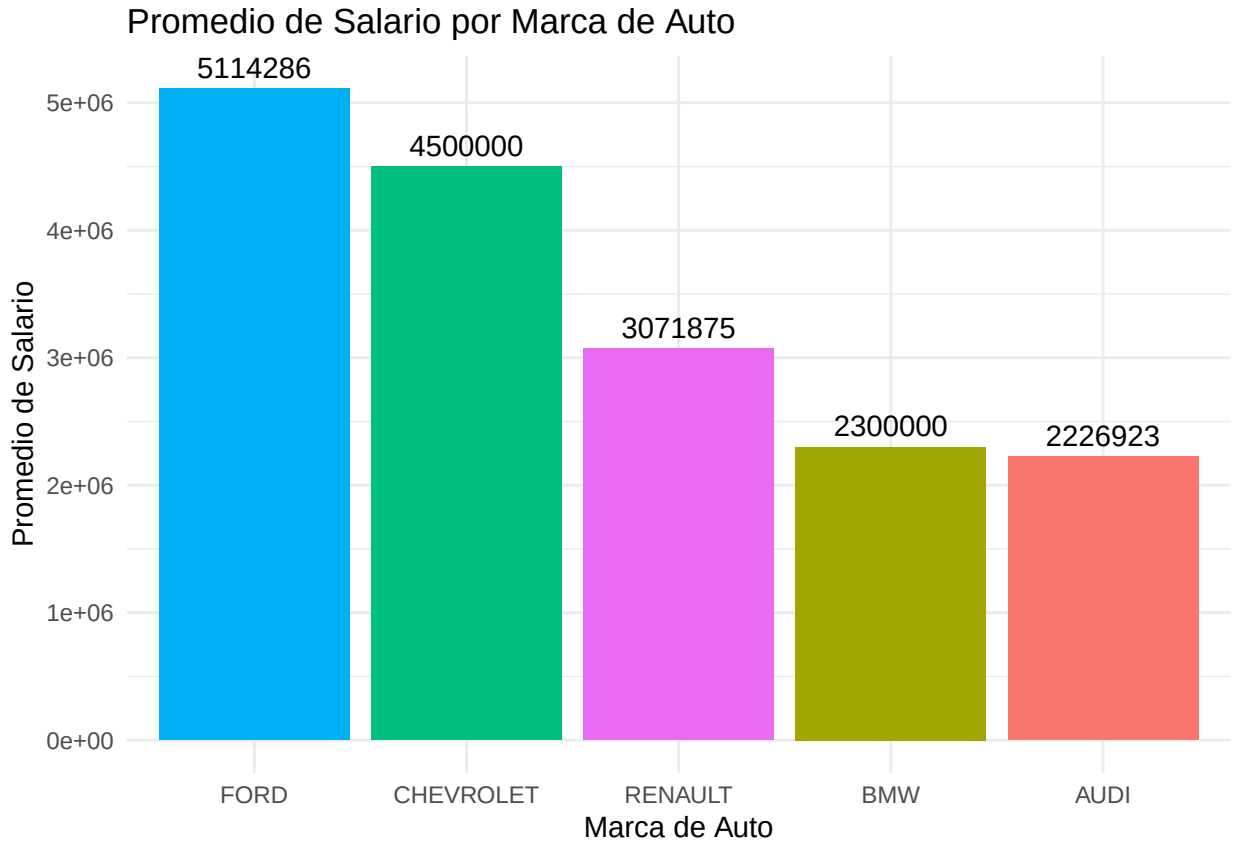
4.4 ¿Cuál es el promedio de salario por cada categoría de la variable "MARCA DE AUTO"?

```r
promedio_salario_por_marca <- data %>%
  group_by(`MARCA DE AUTO`) %>%
  summarise(Promedio_Salario = mean(SALARIO, na.rm = TRUE)) %>%
  arrange(desc(Promedio_Salario))
print(promedio_salario_por_marca)
```

```
## # A tibble: 5 x 2
##   `MARCA DE AUTO` Promedio_Salario
##   <chr>                      <dbl>
## 1 FORD                     5114286.
## 2 CHEVROLET                4500000
## 3 RENAULT                  3071875
## 4 BMW                      2300000
## 5 AUDI                     2226923.
```

```r
ggplot(promedio_salario_por_marca,
       aes(x = reorder(`MARCA DE AUTO`, -Promedio_Salario),
```

```
          y = Promedio_Salario, fill = `MARCA DE AUTO`)) +
geom_bar(stat = "identity") +
geom_text(aes(label = round(Promedio_Salario, 0)), vjust = -0.5) +
labs(
  title = "Promedio de Salario por Marca de Auto",
  x = "Marca de Auto",
  y = "Promedio de Salario"
) +
theme_minimal() +
theme(legend.position = "none")
```



Al analizar el salario promedio de los clientes según la marca de vehículo adquirido, se evidencia una relación interesante entre el nivel de ingresos y las preferencias de compra.

Los clientes que adquirieron vehículos Ford presentan el salario promedio más alto, con aproximadamente 5.1 millones de pesos, seguidos por quienes compraron Chevrolet (4.5 millones). En contraste, marcas comúnmente asociadas al lujo como Audi y BMW registran promedios más bajos, alrededor de 2.3 millones, lo que podría indicar que estos vehículos están siendo adquiridos por clientes que priorizan el prestigio de marca o acceden mediante financiamiento a largo plazo, más que por una elevada capacidad adquisitiva inmediata.

Por su parte, Renault, reconocida por su equilibrio entre accesibilidad y funcionalidad, tiene un salario promedio de 3 millones, posicionándose como una opción atractiva para clientes de ingresos medios.

Esta información resulta clave para diseñar estrategias de mercadeo segmentadas y planes de financiamiento personalizados según el perfil económico del cliente, optimizando así las decisiones comerciales del concesionario.

5 Conclusiones

1. **Calidad y estructura de los datos**: El proceso de limpieza evidenció varios desafíos de calidad en los datos, incluyendo valores faltantes, errores de digitación y variables importadas con el tipo de dato incorrecto. A través de técnicas de normalización, imputación justificada (como el uso de la mediana o asignación aleatoria en distribuciones bimodales) y la conversión de tipos de datos, se logró construir una base confiable para el análisis.

2. **Perfil general del cliente**: El cliente promedio del concesionario es una persona de alrededor de 40 a 50 años, con un nivel educativo alto, un salario promedio cercano a los $3 millones, una estatura promedio de 1.65 m, y en su mayoría sin hijos o con un número reducido. Esta información es clave para adaptar tanto la estrategia de ventas como la atención al cliente.

3. **Preferencias de marca**: Renault es la marca más popular entre los clientes, seguida por Audi, Chevrolet y BMW. Sin embargo, marcas como Ford presentan el salario promedio más alto entre sus compradores, lo que sugiere que ciertos segmentos de alto poder adquisitivo priorizan características distintas a la popularidad, como potencia, seguridad o estatus.

4. **Relación entre nivel educativo e intención de compra**: Se identificó que los clientes con nivel educativo de maestría o doctorado tienden a preferir vehículos de marcas reconocidas. Este segmento también tiene mayor poder adquisitivo, lo que puede orientar campañas de mercadeo personalizadas para profesionales altamente calificados.

5. **Factores adicionales que inciden en la compra**: La presencia de mascotas influye en la elección del vehículo. Clientes con animales de compañía se inclinan por marcas que ofrecen modelos amplios y funcionales. Este tipo de hallazgo permite segmentar el mercado no solo por variables sociodemográficas, sino también por estilo de vida.

6. **Diferencias de género**: Aunque los hombres representan una mayor proporción de compradores, esto puede estar vinculado a factores económicos, culturales y legales. Es importante que el concesionario considere estrategias inclusivas y observe cómo las mujeres pueden representar un mercado potencial aún subestimado.

7. **Importancia del análisis exploratorio**: El uso de histogramas, polígonos de frecuencia y ojivas permitió identificar la distribución de variables clave como edad y estatura. Estos gráficos, junto con los resúmenes estadísticos, revelan patrones útiles para la segmentación de clientes y detección de datos anómalos.

6 Recomendaciones

1. **Mejorar la calidad en la recolección de datos**: Se recomienda implementar formatos digitales o controles de validación al momento de ingresar la información de los clientes, para reducir errores de digitación, valores atípicos y datos faltantes. Esto no solo mejora la calidad del análisis, sino que también optimiza futuras estrategias de CRM (Customer Relationship Management).

2. **Segmentación de campañas de mercadeo**:

- Crear campañas dirigidas a clientes con nivel educativo alto (maestría y doctorado), destacando características de vehículos de gama media-alta como tecnología, eficiencia y seguridad.
- Diseñar estrategias específicas para clientes con mascotas, resaltando modelos con mayor espacio interior, sistemas de ventilación trasera, accesorios para animales y facilidad de limpieza.

3. **Explorar oportunidades de mercado entre mujeres**: Aunque actualmente hay más hombres compradores, el segmento femenino está casi a la par. Se recomienda realizar campañas específicas dirigidas a mujeres, resaltando atributos como diseño, seguridad, economía de combustible y facilidad de conducción.

4. **Ofrecer planes de financiamiento diferenciados**: Dado que marcas como Audi y BMW fueron adquiridas por clientes con salarios no necesariamente más altos, es recomendable ofrecer financiación flexible o convenios con entidades bancarias para atraer a este perfil aspiracional que valora el prestigio de la marca.

5. **Aprovechar el perfil del cliente promedio**: Teniendo en cuenta que el cliente promedio tiene entre 40 y 50 años, sin muchos hijos, con un salario medio y educación superior, el concesionario puede priorizar modelos compactos, seguros y con buena relación costo-beneficio. También es ideal considerar servicios complementarios como mantenimiento preventivo, garantías extendidas o seguros personalizados.

6. **Analizar de forma cruzada otras variables**: Se sugiere explorar relaciones más profundas entre variables como salario, número de hijos, nivel educativo y marca del vehículo para diseñar un perfil psicográfico más robusto. Esto podría incluir métodos de análisis multivariado o clustering en futuros estudios.

7. **Continuar con análisis periódicos**: La periodicidad en el análisis de los datos de clientes puede permitir identificar cambios en el comportamiento del consumidor, nuevas tendencias y oportunidades de mercado. Se recomienda establecer reportes trimestrales o semestrales para una toma de decisiones basada en datos actualizados.

7 Recomendaciones Comerciales

A partir del análisis realizado sobre el perfil de los clientes del concesionario, se proponen las siguientes recomendaciones orientadas a mejorar la estrategias de ventas, segmentación y personalización de la oferta:

1. **Ofertas dirigidas a profesionales con posgrado**: Se identificó que los clientes con niveles educativos altos (maestría y doctorado) tienden a adquirir vehículos de marcas reconocidas como Audi, BMW y Renault. Este grupo también presenta un salario promedio elevado, lo que indica un mayor poder adquisitivo. Por lo tanto se recomienda:

    - Diseñar planes de financiamiento personalizados para este segmento
    - Ofrecer beneficios adicionales como mantenimientos incluidos o garantías extendidas
    - Realizar campañas dirigidas a profesionales universitarios y profesionales con posgrados, especialmente a partir de los 30 años.

2. **Campañas orientadas a clientes con mascotas**: El 50% de los clientes tienen mascotas y tienden a preferir vehículos con mayor espacio interior. Se suguiere:

    - Promocionar modelos familiares, SUV o con baúl amplio a este grupo
    - Incluir accesorios especiales para transporte de mascotas como parte de la oferta
    - Utilizar esta variable para segmentar promociones en canales digitales

3. **Estrategias por grupo etario**: La mayoría de los clientes se concentran en dos franjas: adultos entre 40 y 60 años, y jóvenes entre 25 y 35 años. A cada grupo podrían dirigirse diferentes estrategias:

    - A los adultos, ofrecer vehículos con confort y reputación de marca (Renault, Audi)
    - A los jóvenes, destacar vehículos más accesibles y compactos, o promover planes de leasing

4. **Optimización del portafolio de marcas**: Renault y Audi fueron las marcas más vendidas entre varios subgrupos analizados. Esto sugiere mantener y reforzar su disponibilidad. Sin embargo, marcas como Ford presentan menor volumen de ventas y podrían evaluarse para estrategias específicas como:

    - Ofertas de cierre de inventario
    - Paquetes con mantenimiento gratuito para aumentar su atractivo

5. **Uso estratégico del número de hijos como variable comercial**: Dado que la mayoría de clientes tienen entre 0 y 2 hijos, los vehículos compactos o sedanes familiares tienen alta relevancia. También se recomienda:

    - Evaluar la demanda potencial de vehículos más grandes en casos de familias numerosas.
    - Ofrecer test drives o beneficios específicos para quienes buscan autos familiares