ProyectoUnidad3

2024-05-27

Formulación de Preguntas e Hipótesis

Preguntas de Investigación

- ¿Cómo afecta el tipo de dispositivo al número de sesiones de los usuarios?
- ¿Existe una relación entre los kilómetros conducidos y el número de sesiones?
- ¿Los días de actividad afectan la retención de los usuarios

Hipótesis Relacionadas

- Usuarios con Android tienen más sesiones que aquellos con iPhone
- Usuarios que conducen más kilómetros tienen sesiones más frecuentes.
- Un mayor número de días de actividad conduce a una menor tasa de abandono.

Importación de datos

Este dataset fue obtenido desde Kaggle (https://www.kaggle.com/datasets/raminhuseyn/wase-navigation-app-dataset?select=waze_app_dataset.csv) , luego fue re-subido a un repositorio de github para poder ser utilizado desde una URL directamente.

```
library(readr)
url <- "https://raw.githubusercontent.com/ShinichiKD/analisisExploratorio/main/waze_app_dataset.csv"
waze_app_dataset <- read_csv(url,
    col_names = TRUE, col_types = cols(label = col_factor(levels = c("retained",
        "churned")), device = col_factor(levels = c("Android",
        "iPhone"))), skip = 0)
```

```
## ‘curl‘ package not installed, falling back to using ‘url()‘
```

```
head(waze_app_dataset)
```

```
## # A tibble: 6 x 13
##      ID label    sessions drives total_sessions n_days_after_onboarding
##   <dbl> <fct>       <dbl>  <dbl>          <dbl>                   <dbl>
## 1     0 retained      283    226          297.                     2276
## 2     1 retained      133    107          327.                     1225
## 3     2 retained      114     95          136.                     2651
## 4     3 retained       49     40           67.6                      15
## 5     4 retained       84     68          168.                     1562
```

```
## 6     5 retained      113    103          280.                      2637
## # i 7 more variables: total_navigations_fav1 <dbl>,
## #   total_navigations_fav2 <dbl>, driven_km_drives <dbl>,
## #   duration_minutes_drives <dbl>, activity_days <dbl>, driving_days <dbl>,
## #   device <fct>
```

```
str(waze_app_dataset)
```

```
## spc_tbl_ [14,999 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ID                      : num [1:14999] 0 1 2 3 4 5 6 7 8 9 ...
##  $ label                   : Factor w/ 2 levels "retained","churned": 1 1 1 1 1 1 1 1 1 2 ...
##  $ sessions                : num [1:14999] 283 133 114 49 84 113 3 39 57 84 ...
##  $ drives                  : num [1:14999] 226 107 95 40 68 103 2 35 46 68 ...
##  $ total_sessions          : num [1:14999] 296.7 326.9 135.5 67.6 168.2 ...
##  $ n_days_after_onboarding: num [1:14999] 2276 1225 2651 15 1562 ...
##  $ total_navigations_fav1 : num [1:14999] 208 19 0 322 166 0 185 0 0 72 ...
##  $ total_navigations_fav2 : num [1:14999] 0 64 0 7 5 0 18 0 26 0 ...
##  $ driven_km_drives       : num [1:14999] 2629 13716 3059 914 3950 ...
##  $ duration_minutes_drives: num [1:14999] 1986 3160 1611 587 1220 ...
##  $ activity_days          : num [1:14999] 28 13 14 7 27 15 28 22 25 7 ...
##  $ driving_days           : num [1:14999] 19 11 8 3 18 11 23 20 20 3 ...
##  $ device                  : Factor w/ 2 levels "Android","iPhone": 1 2 1 2 1 2 2 2 1 2 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ID = col_double(),
##   ..   label = col_factor(levels = c("retained", "churned"), ordered = FALSE, include_na = FALSE),
##   ..   sessions = col_double(),
##   ..   drives = col_double(),
##   ..   total_sessions = col_double(),
##   ..   n_days_after_onboarding = col_double(),
##   ..   total_navigations_fav1 = col_double(),
##   ..   total_navigations_fav2 = col_double(),
##   ..   driven_km_drives = col_double(),
##   ..   duration_minutes_drives = col_double(),
##   ..   activity_days = col_double(),
##   ..   driving_days = col_double(),
##   ..   device = col_factor(levels = c("Android", "iPhone"), ordered = FALSE, include_na = FALSE)
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Acerca del dataset

El conjunto de datos ofrece una visión completa de las interacciones de los usuarios dentro de la aplicación de navegación Waze, crucial para comprender y mitigar la pérdida de usuarios.

Variables presentes

- **ID** Identificador único para cada usuario.
- **label** Etiqueta que indica el estado de abandono de usuarios (p. ej., abandonado, retenido).
- **sessions** Número de sesiones registradas por el usuario.
- **drives** Número de unidades completadas por el usuario.

- **total_sessions** Número total de sesiones registradas para el usuario.
- **n_days_after_onboarding** Número de días desde la incorporación del usuario.
- **total_navigations_fav1** Número total de navegaciones por ruta favorita 1.
- **total_navigations_fav2** Número total de navegaciones por ruta favorita 2.
- **driven_km_drives** Distancia total recorrida por el usuario en kilómetros.
- **duration_minutes_drives** Duración total de los recorridos en minutos.
- **activity_days** Número de días con actividad de usuario registrada.
- **driving_days** Número de días con actividad de conducción registrada.
- **device** Dispositivo utilizado por el usuario para la navegación (p. ej., teléfono inteligente, tableta).

Limpieza de Datos

Datos faltantes

```
totalFaltantes<-sum(is.na(waze_app_dataset))

numero_de_filas <- nrow(waze_app_dataset)
porcentaje_faltantes <- (totalFaltantes / numero_de_filas) * 100
summary(waze_app_dataset)
```

```
##        ID             label          sessions          drives
##  Min.   :    0   retained:11763   Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 3750   churned : 2536   1st Qu.: 23.00   1st Qu.: 20.00
##  Median : 7499   NA's    :  700   Median : 56.00   Median : 48.00
##  Mean   : 7499                    Mean   : 80.63   Mean   : 67.28
##  3rd Qu.:11248                    3rd Qu.:112.00   3rd Qu.: 93.00
##  Max.   :14998                    Max.   :743.00   Max.   :596.00
##  total_sessions      n_days_after_onboarding total_navigations_fav1
##  Min.   :   0.2202   Min.   :   4            Min.   :   0.0
##  1st Qu.:  90.6612   1st Qu.: 878            1st Qu.:   9.0
##  Median : 159.5681   Median :1741            Median :  71.0
##  Mean   : 189.9644   Mean   :1750            Mean   : 121.6
##  3rd Qu.: 254.1923   3rd Qu.:2624            3rd Qu.: 178.0
##  Max.   :1216.1546   Max.   :3500            Max.   :1236.0
##  total_navigations_fav2 driven_km_drives   duration_minutes_drives
##  Min.   :  0.00         Min.   :   60.44   Min.   :   18.28
##  1st Qu.:  0.00         1st Qu.: 2212.60   1st Qu.:  836.00
##  Median :  9.00         Median : 3493.86   Median : 1478.25
##  Mean   : 29.67         Mean   : 4039.34   Mean   : 1860.98
##  3rd Qu.: 43.00         3rd Qu.: 5289.86   3rd Qu.: 2464.36
##  Max.   :415.00         Max.   :21183.40   Max.   :15851.73
##  activity_days    driving_days       device
##  Min.   : 0.00   Min.   : 0.00   Android:5327
##  1st Qu.: 8.00   1st Qu.: 5.00   iPhone :9672
##  Median :16.00   Median :12.00
##  Mean   :15.54   Mean   :12.18
##  3rd Qu.:23.00   3rd Qu.:19.00
##  Max.   :31.00   Max.   :30.00
```

Manejo de datos faltantes

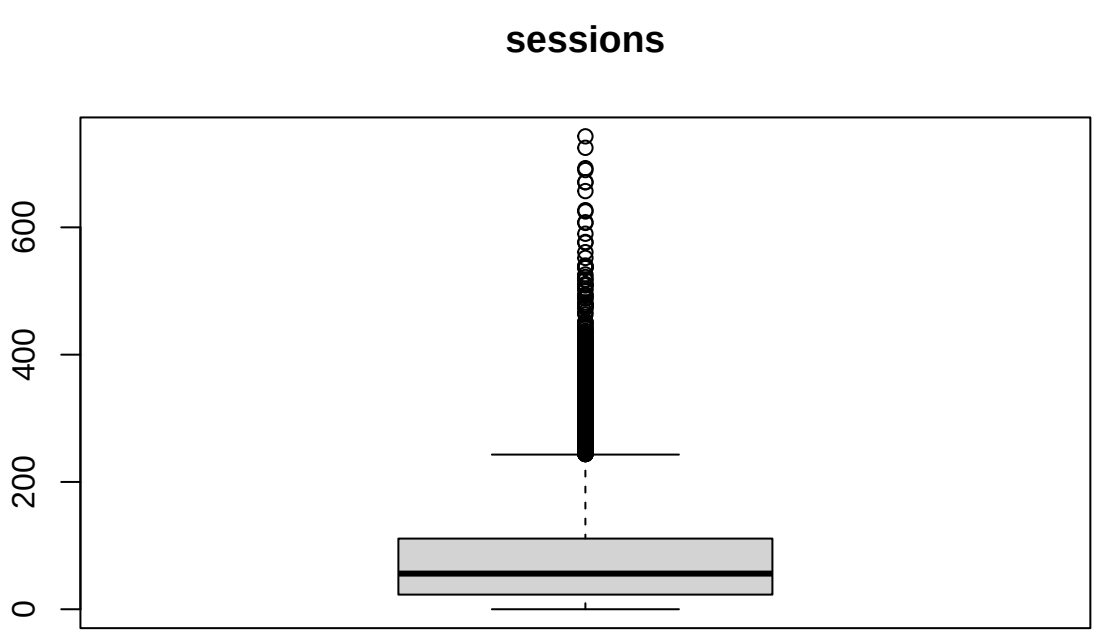
En el análisis se encontró que tenemos un total de 700 de datos faltantes, además observando el resumen , podemos ver que todos corresponden a la columna "label", obteniendo un 4.6669778% de datos faltantes en el dataset. Esta cantidad de datos no es significativa , por lo tanto estas filas seran eliminadas.

```
waze_app_dataset_limpios <- na.omit(waze_app_dataset)
summary(waze_app_dataset_limpios)
```

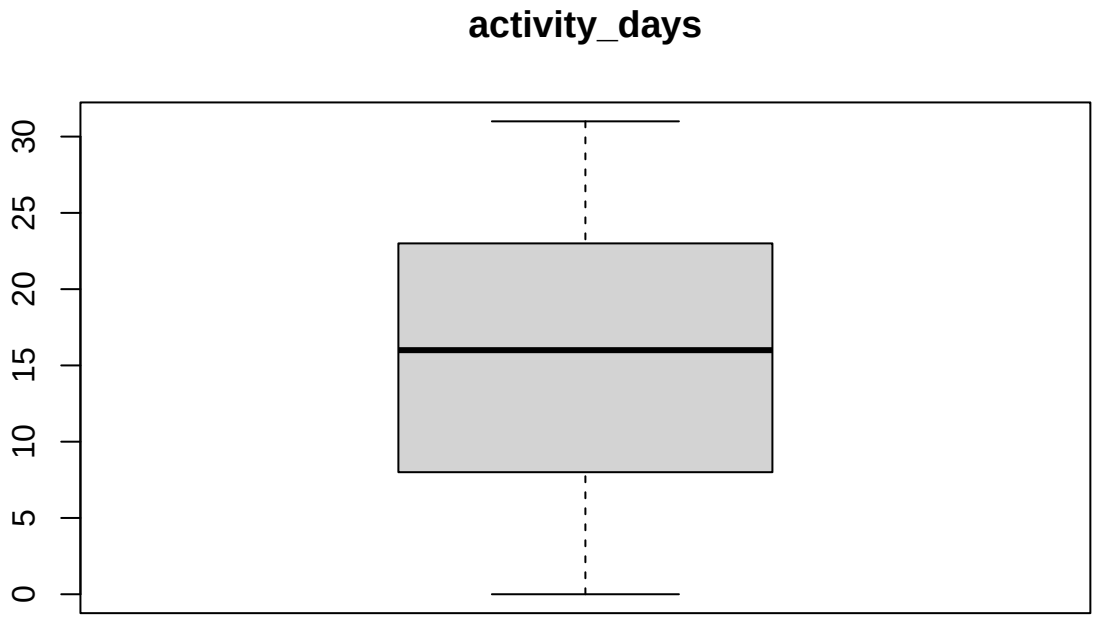
```
##        ID             label          sessions          drives
##  Min.   :    0   retained:11763   Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 3750   churned : 2536   1st Qu.: 23.00   1st Qu.: 20.00
##  Median : 7504                    Median : 56.00   Median : 48.00
##  Mean   : 7504                    Mean   : 80.62   Mean   : 67.26
##  3rd Qu.:11258                    3rd Qu.:111.00   3rd Qu.: 93.00
##  Max.   :14998                    Max.   :743.00   Max.   :596.00
##  total_sessions      n_days_after_onboarding total_navigations_fav1
##  Min.   :   0.2202   Min.   :   4.0          Min.   :   0.0
##  1st Qu.:  90.4577   1st Qu.: 878.5          1st Qu.:  10.0
##  Median : 158.7186   Median :1749.0          Median :  71.0
##  Mean   : 189.5474   Mean   :1751.8          Mean   : 121.7
##  3rd Qu.: 253.5404   3rd Qu.:2627.5          3rd Qu.: 178.0
##  Max.   :1216.1546   Max.   :3500.0          Max.   :1236.0
##  total_navigations_fav2 driven_km_drives   duration_minutes_drives
##  Min.   :  0.00         Min.   :   60.44   Min.   :   18.28
##  1st Qu.:  0.00         1st Qu.: 2217.32   1st Qu.:  840.18
##  Median :  9.00         Median : 3496.55   Median : 1479.39
##  Mean   : 29.64         Mean   : 4044.40   Mean   : 1864.20
##  3rd Qu.: 43.00         3rd Qu.: 5299.97   3rd Qu.: 2466.93
##  Max.   :415.00         Max.   :21183.40   Max.   :15851.73
##  activity_days    driving_days       device
##  Min.   : 0.00   Min.   : 0.00   Android:5074
##  1st Qu.: 8.00   1st Qu.: 5.00   iPhone :9225
##  Median :16.00   Median :12.00
##  Mean   :15.54   Mean   :12.18
##  3rd Qu.:23.00   3rd Qu.:19.00
##  Max.   :31.00   Max.   :30.00
```

Detectar datos atipicos

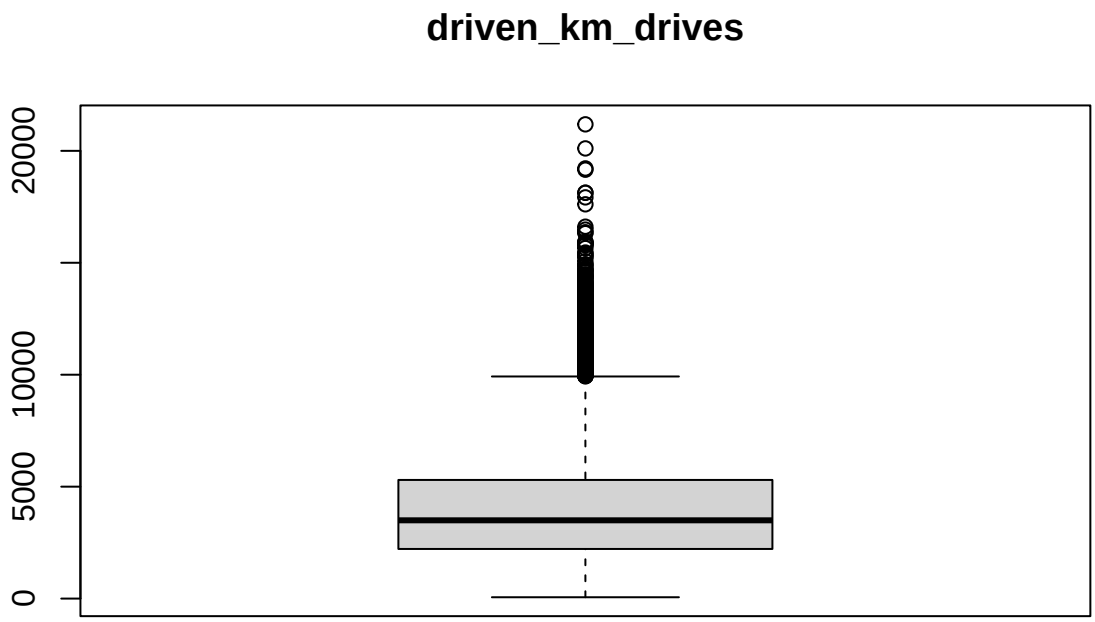
```
boxplot(waze_app_dataset_limpios$sessions, main="sessions")
```




```
boxplot(waze_app_dataset_limpios$activity_days, main="activity_days")
```

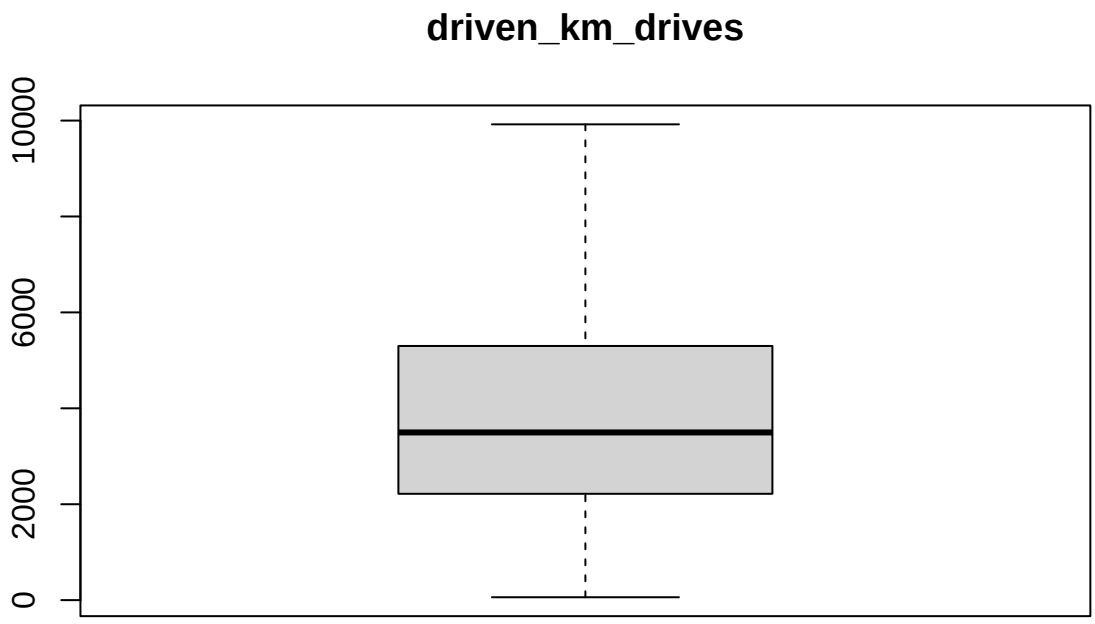



```
boxplot(waze_app_dataset_limpios$driven_km_drives, main="driven_km_drives")
```

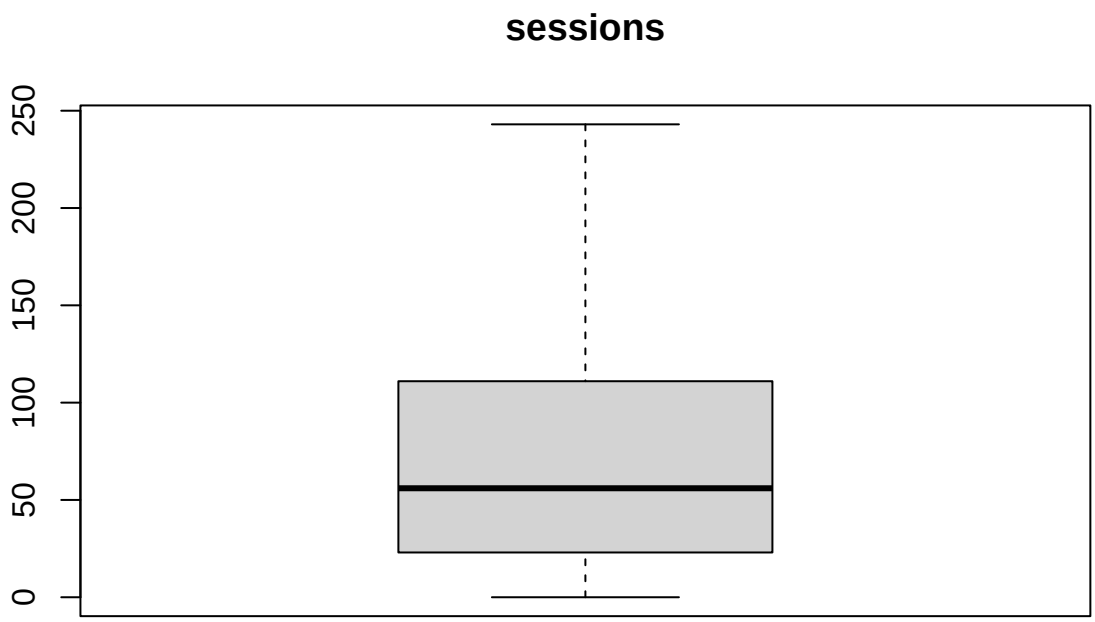


Tratar datos atipicos

```
adjust_outliers <- function(data, column_name) {

    x <- data[[column_name]]
    qnt <- quantile(x, probs=c(.25, .75), na.rm = TRUE)
    caps <- quantile(x, probs=c(.05, .95), na.rm = TRUE)
    H <- 1.5 * IQR(x, na.rm = TRUE)
    x[x < (qnt[1] - H)] <- caps[1]
    x[x > (qnt[2] + H)] <- caps[2]
    data[[column_name]] <- x
    return(data)
}

# Ajustar outliers en la columna 'driven_km_drives'
waze_app_dataset_limpios <- adjust_outliers(waze_app_dataset_limpios, "driven_km_drives")
waze_app_dataset_limpios <- adjust_outliers(waze_app_dataset_limpios, "sessions")

boxplot(waze_app_dataset_limpios$driven_km_drives, main="driven_km_drives")
```




```
boxplot(waze_app_dataset_limpios$sessions , main="sessions")
```



Escalado de los 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(tidyr)
waze_app_dataset_limpios
```

```
## # A tibble: 14,299 x 13
##       ID label    sessions drives total_sessions n_days_after_onboarding
##    <dbl> <fct>       <dbl>  <dbl>          <dbl>                    <dbl>
##  1     0 retained      243    226           297.                     2276
##  2     1 retained      133    107           327.                     1225
##  3     2 retained      114     95           136.                     2651
##  4     3 retained       49     40            67.6                      15
```

```
##  5     4 retained       84     68          168.                      1562
##  6     5 retained      113    103          280.                      2637
##  7     6 retained        3      2          237.                       360
##  8     7 retained       39     35          176.                      2999
##  9     8 retained       57     46          184.                       424
## 10     9 churned        84     68          245.                      2997
## # i 14,289 more rows
## # i 7 more variables: total_navigations_fav1 <dbl>,
## #   total_navigations_fav2 <dbl>, driven_km_drives <dbl>,
## #   duration_minutes_drives <dbl>, activity_days <dbl>, driving_days <dbl>,
## #   device <fct>
```

```
waze_app_dataset_limpios$driven_km_drives_scaled <- scale(waze_app_dataset_limpios$driven_km_drives)
waze_app_dataset_limpios
```

```
## # A tibble: 14,299 x 14
##       ID label    sessions drives total_sessions n_days_after_onboarding
##    <dbl> <fct>       <dbl>  <dbl>          <dbl>                   <dbl>
##  1     0 retained      243    226          297.                     2276
##  2     1 retained      133    107          327.                     1225
##  3     2 retained      114     95          136.                     2651
##  4     3 retained       49     40           67.6                      15
##  5     4 retained       84     68          168.                     1562
##  6     5 retained      113    103          280.                     2637
##  7     6 retained        3      2          237.                      360
##  8     7 retained       39     35          176.                     2999
##  9     8 retained       57     46          184.                      424
## 10     9 churned        84     68          245.                     2997
## # i 14,289 more rows
## # i 8 more variables: total_navigations_fav1 <dbl>,
## #   total_navigations_fav2 <dbl>, driven_km_drives <dbl>,
## #   duration_minutes_drives <dbl>, activity_days <dbl>, driving_days <dbl>,
## #   device <fct>, driven_km_drives_scaled <dbl[,1]>
```

```
# Medidas de tendencia central y dispersión para variables numéricas
waze_app_dataset_limpios %>%
  summarise(across(where(is.numeric), list(media = ~mean(.), mediana = ~median(.), desviacion = ~sd(.))))
```

```
## # A tibble: 1 x 36
##   ID_media ID_mediana ID_desviacion sessions_media sessions_mediana
##      <dbl>      <dbl>         <dbl>          <dbl>            <dbl>
## 1    7504.       7504         4331.           76.5               56
## # i 31 more variables: sessions_desviacion <dbl>, drives_media <dbl>,
## #   drives_mediana <dbl>, drives_desviacion <dbl>, total_sessions_media <dbl>,
## #   total_sessions_mediana <dbl>, total_sessions_desviacion <dbl>,
## #   n_days_after_onboarding_media <dbl>, n_days_after_onboarding_mediana <dbl>,
## #   n_days_after_onboarding_desviacion <dbl>,
## #   total_navigations_fav1_media <dbl>, total_navigations_fav1_mediana <dbl>,
## #   total_navigations_fav1_desviacion <dbl>, ...
```

Análisis descriptivo

Variables numericas

```
library(dplyr)
library(tidyr)

# Medidas de tendencia central y dispersión para variables numéricas
waze_app_dataset_limpios %>%
  summarise(across(where(is.numeric), list(media = ~mean(.), mediana = ~median(.), desviacion = ~sd(.))))
```

```
## # A tibble: 1 x 36
##   ID_media ID_mediana ID_desviacion sessions_media sessions_mediana
##      <dbl>      <dbl>         <dbl>          <dbl>            <dbl>
## 1    7504.       7504         4331.           76.5               56
## # i 31 more variables: sessions_desviacion <dbl>, drives_media <dbl>,
## #   drives_mediana <dbl>, drives_desviacion <dbl>, total_sessions_media <dbl>,
## #   total_sessions_mediana <dbl>, total_sessions_desviacion <dbl>,
## #   n_days_after_onboarding_media <dbl>, n_days_after_onboarding_mediana <dbl>,
## #   n_days_after_onboarding_desviacion <dbl>,
## #   total_navigations_fav1_media <dbl>, total_navigations_fav1_mediana <dbl>,
## #   total_navigations_fav1_desviacion <dbl>, ...
```

Variables categóricas

```
library(dplyr)
library(tidyr)


# Análisis descriptivo para variables categóricas
summary(factor(waze_app_dataset_limpios$device))
```

```
## Android  iPhone
##    5074    9225
```

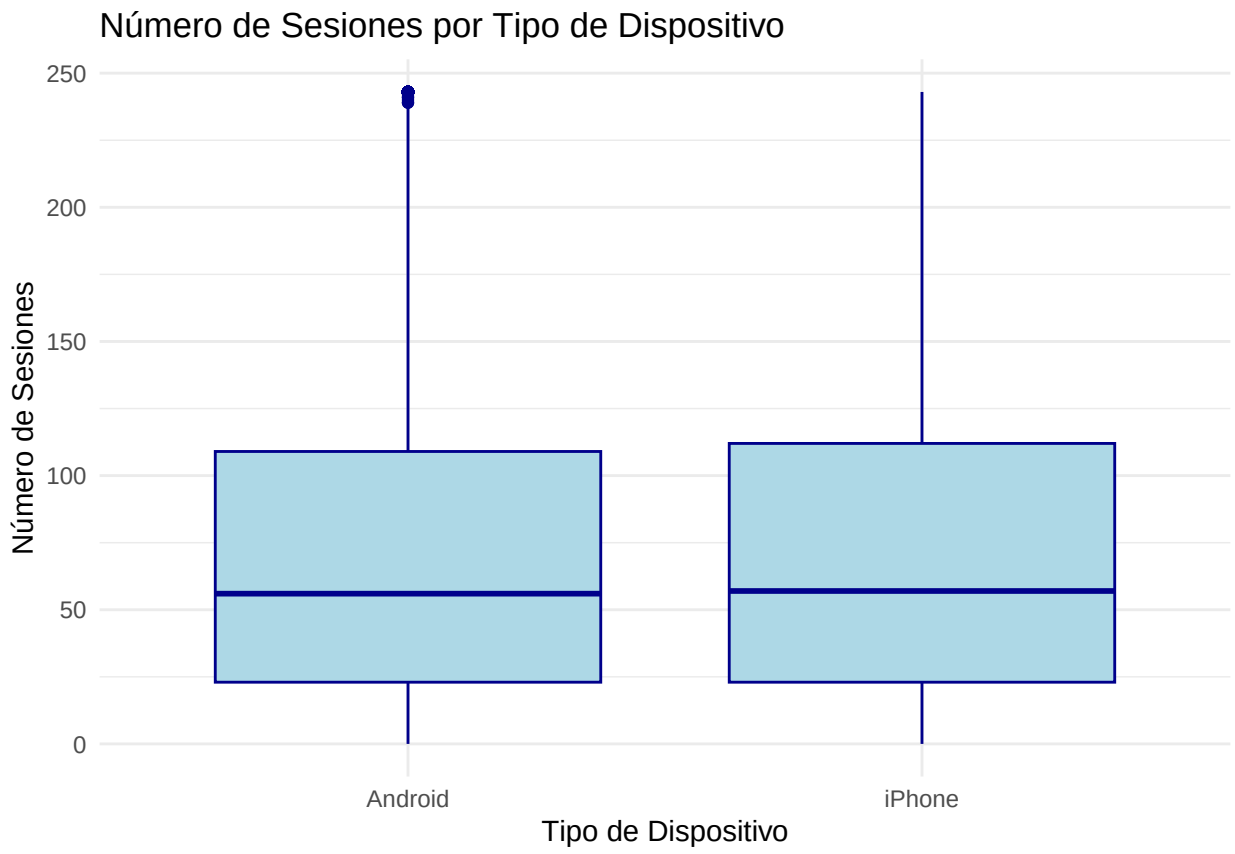
```
summary(factor(waze_app_dataset_limpios$label))
```

```
## retained  churned
##    11763     2536
```

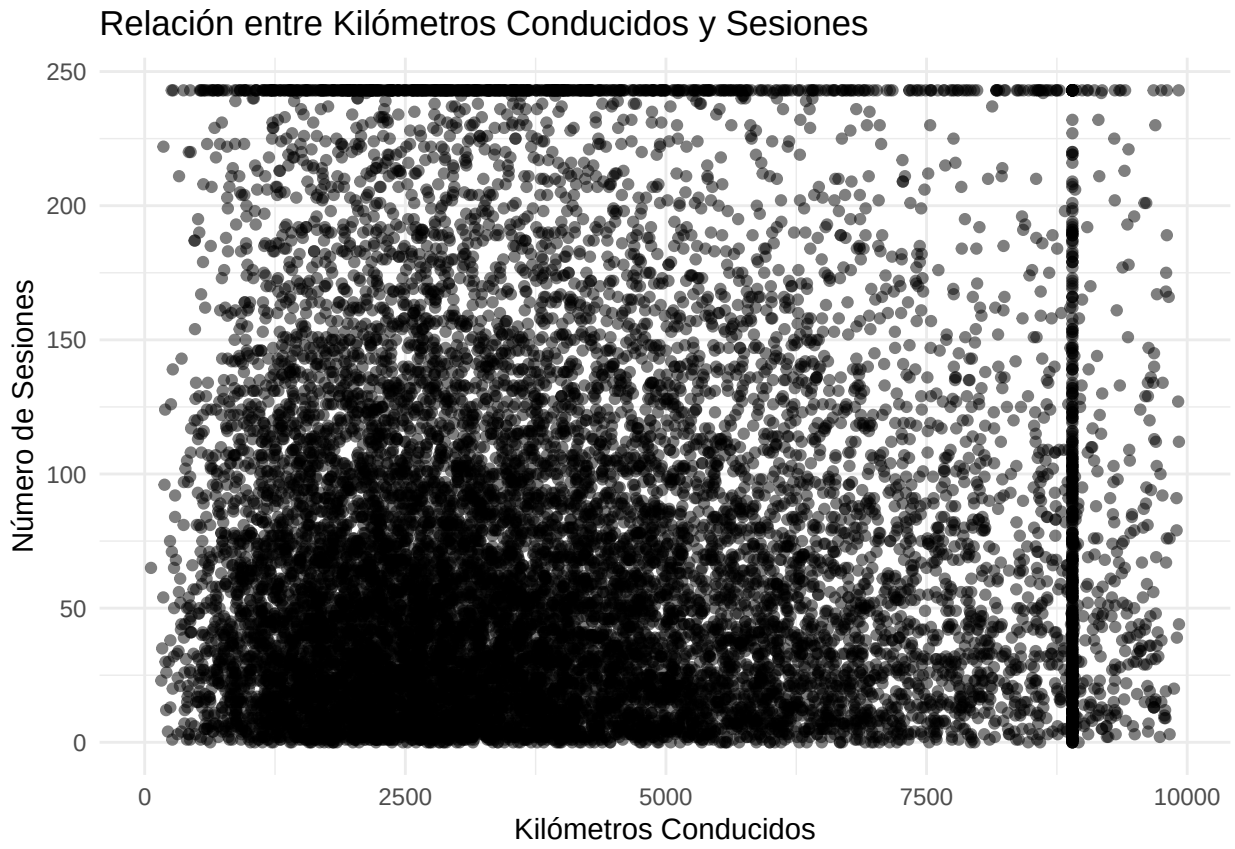
Visualización de datos

```
library(ggplot2)


ggplot(waze_app_dataset_limpios, aes(x=device, y=sessions)) +
```

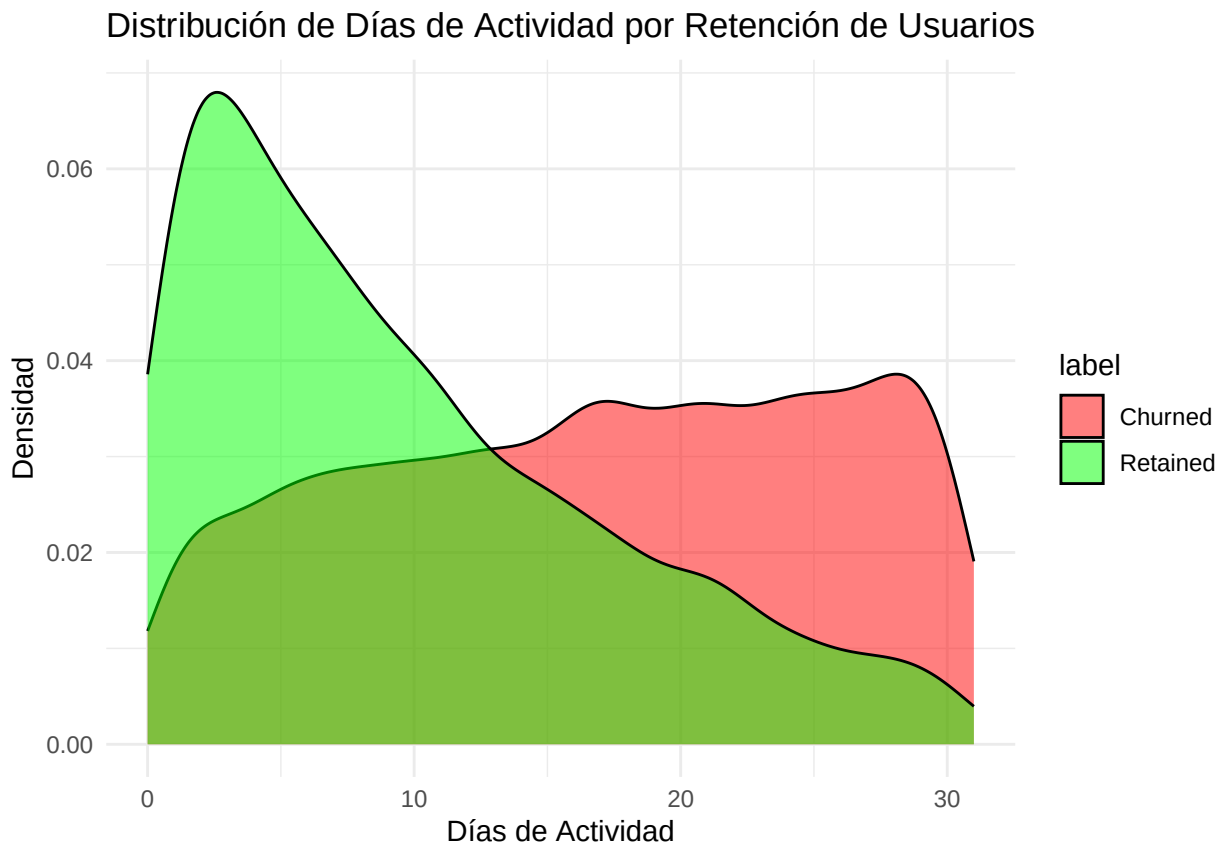
```
  geom_boxplot(fill="lightblue", color="darkblue") +
  labs(title="Número de Sesiones por Tipo de Dispositivo",
       x="Tipo de Dispositivo",
       y="Número de Sesiones") +
  theme_minimal()
```




```
ggplot(waze_app_dataset_limpios, aes(x=driven_km_drives, y=sessions)) +
  geom_point(alpha=0.5) +
  labs(title="Relación entre Kilómetros Conducidos y Sesiones",
       x="Kilómetros Conducidos",
       y="Número de Sesiones") +
  theme_minimal()
```




```
ggplot(waze_app_dataset_limpios, aes(x=activity_days, fill=label)) +
  geom_density(alpha=0.5) +
  labs(title="Distribución de Días de Actividad por Retención de Usuarios",
       x="Días de Actividad",
       y="Densidad") +
  scale_fill_manual(values=c("red", "green"), labels=c("Churned", "Retained")) +
  theme_minimal()
```



Interpretaciones y conclusiones

¿Cómo afecta el tipo de dispositivo al número de sesiones de los usuarios?

El boxplot muestra que la mediana del número de sesiones para usuarios de iPhone parece ser ligeramente superior a la de los usuarios de Android, aunque ambos dispositivos muestran una amplia variabilidad en el número de sesiones. La dispersión en ambos dispositivos indica que, aunque la tendencia central es similar, existen usuarios en ambos extremos que usan la aplicación con frecuencias muy diferentes.

La hipótesis inicial de que los usuarios de Android tienen más sesiones que aquellos con iPhone no se sostiene completamente; los resultados indican que los usuarios de iPhone podrían tener ligeramente más sesiones en promedio.

¿Existe una relación entre los kilómetros conducidos y el número de sesiones?

Se puede observar una distribución amplia sin una tendencia clara, indicando que no hay una relación lineal fuerte entre los kilómetros conducidos y el número de sesiones. Algunos usuarios con pocos kilómetros conducidos tienen un alto número de sesiones y viceversa, lo que podría indicar que otros factores (como el tipo de uso o necesidades individuales) influyen más en el número de sesiones que la distancia conducida sola.

La hipótesis de que los usuarios que conducen más kilómetros tienen sesiones más frecuentes no se apoya fuertemente con estos datos. Sería bueno explorar otros factores que podrían influir en el número de sesiones.