Práctica 2

Lenin Torres

1/1/2020

Contents

1 Descripción del dataset

El conjunto de datos seleccionado se obtuvo desde https://www.kaggle.com/roshansharma/online-shoppers-intention y de acuerdo a la descripciòn del mismo, contiene datos de 12330 sesiones de usuarios que realizan compras en línea. Este conjunto de datos se selecciona debido a que será interesante explorar modelos que permitan detectar cuales factores contribuyen a que la intención de compra de usuarios de tiendas en línea efectivamente realicen una compra.

Los atributos del conjunto de datos son los siguientes:

- Los siguientes atributos indican el número de páginas y el tiempo total que el usuario permaneció en tres categorías de página: administrativa, informativa y de producto relacionado.
  - Administrative,
  - Administrative Duration,
  - Informational,
  - Informational Duration,
  - Product Related
  - Product Related Duration

- Atributos generados con Google Analytics:
  - Bunce rate. Porcentaje de visitantes que ingresan al sitio usando una página pero que lo abandonan sin realizar ninguna acción adicional
  - Exit rate: De todas las visitas a la página, el porcentje en el que fue la última en la sesión del usuario.
  - Page value: valor promedio de una página visitada antes de haber completado una transaccion.
- Atributos relacionados con el navegador empleado por el usuario:

  - OperatingSystems
  - Browser
- Traffict type
- Visitor Type
- Special day. Indica la cercanía de la visita a una fecha comercial importante, como el día de San Valentín
- Weekend. Indica si la sesión de usuario se realizo en en fin de semana o no
- Revenue. Indica si la visita finaliza o no en una transacciòn

El objetivo que se persigue en este trabajo es predecir el valor del atributo *Revenue* mediante la creación de un modelo.

2 Integración y selección de los datos de interés a analizar.

Para el caso de esta práctica, se opta por utilizar todos los atributos relacionados para realzar una predicción de la variable *Revenue*, se omiten los atributos relacionados con el navegador empleado

3 Limpieza de los datos.

Iniciamos leyendo el archivo y presentando un resumen

```
datos <- read.csv(file = "online_shoppers_intention.csv",sep = ",", dec = ".")
str(datos)
```

```
## 'data.frame':    12330 obs. of  18 variables:
##  $ Administrative         : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Administrative_Duration: num  0 0 -1 0 0 0 -1 -1 0 0 ...
##  $ Informational          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational_Duration : num  0 0 -1 0 0 0 -1 -1 0 0 ...
##  $ ProductRelated         : int  1 2 1 2 10 19 1 1 2 3 ...
##  $ ProductRelated_Duration: num  0 64 -1 2.67 627.5 ...
##  $ BounceRates            : num  0.2 0 0.2 0.05 0.02 ...
##  $ ExitRates              : num  0.2 0.1 0.2 0.14 0.05 ...
##  $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SpecialDay             : num  0 0 0 0 0 0 0.4 0 0.8 0.4 ...
##  $ Month                  : Factor w/ 10 levels "Aug","Dec","Feb",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ OperatingSystems       : int  1 2 4 3 3 2 2 1 2 2 ...
##  $ Browser                : int  1 2 1 2 3 2 4 2 2 4 ...
##  $ Region                 : int  1 1 9 2 1 1 3 1 2 1 ...
##  $ TrafficType            : int  1 2 3 4 4 3 3 5 3 2 ...
##  $ VisitorType            : Factor w/ 3 levels "New_Visitor",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Weekend                : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ Revenue                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

Para facilitar la selecciòn de atributos, se separan en dos estructuras

```
atribNumericos = c("Administrative","Administrative_Duration","Informational","Informational_Duration",
atribCategoricos = c("Revenue")
datos$Revenue = as.factor(datos$Revenue)
datos = datos[,names(datos) %in% c(atribNumericos,atribCategoricos)]
```

Se revisa nuevamente el conjunto de datos.

```
str(datos)
```

```
## 'data.frame':    12330 obs. of  10 variables:
##  $ Administrative         : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Administrative_Duration: num  0 0 -1 0 0 0 -1 -1 0 0 ...
##  $ Informational          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational_Duration : num  0 0 -1 0 0 0 -1 -1 0 0 ...
##  $ ProductRelated         : int  1 2 1 2 10 19 1 1 2 3 ...
##  $ ProductRelated_Duration: num  0 64 -1 2.67 627.5 ...
##  $ BounceRates            : num  0.2 0 0.2 0.05 0.02 ...
##  $ ExitRates              : num  0.2 0.1 0.2 0.14 0.05 ...
##  $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Revenue                : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 1 1 1 1 1 ...
```

3.1 Datos perdídos

Primeramente comprobamos si en la lectura algunos datos se marcaron como NA:

```
colSums(is.na(datos))
```

```
##          Administrative Administrative_Duration           Informational
##                      14                      14                      14
##  Informational_Duration          ProductRelated ProductRelated_Duration
##                      14                      14                      14
##             BounceRates              ExitRates             PageValues
##                      14                      14                       0
##                 Revenue
##                       0
```

Se puede observar que existen exactamente 14 ocurrencias en las que se cargaron valores perdídos, esto representa un 0.11% de las observaciones. Examinando algunos ejemplos de estas observaciones, se puede observar que los valores perdídos provienen de las mismas 14 filas:

```
datos[is.na(datos$Administrative),]
```

```
##      Administrative Administrative_Duration Informational
## 1066             NA                      NA            NA
## 1133             NA                      NA            NA
## 1134             NA                      NA            NA
## 1135             NA                      NA            NA
## 1136             NA                      NA            NA
## 1137             NA                      NA            NA
## 1474             NA                      NA            NA
## 1475             NA                      NA            NA
## 1476             NA                      NA            NA
## 1477             NA                      NA            NA
## 2038             NA                      NA            NA
## 2039             NA                      NA            NA
## 2040             NA                      NA            NA
## 2754             NA                      NA            NA
##      Informational_Duration ProductRelated ProductRelated_Duration
## 1066                     NA             NA                      NA
## 1133                     NA             NA                      NA
## 1134                     NA             NA                      NA
## 1135                     NA             NA                      NA
```

```
## 1136                    NA             NA                      NA
## 1137                    NA             NA                      NA
## 1474                    NA             NA                      NA
## 1475                    NA             NA                      NA
## 1476                    NA             NA                      NA
## 1477                    NA             NA                      NA
## 2038                    NA             NA                      NA
## 2039                    NA             NA                      NA
## 2040                    NA             NA                      NA
## 2754                    NA             NA                      NA
##      BounceRates ExitRates PageValues Revenue
## 1066          NA        NA          0   FALSE
## 1133          NA        NA          0   FALSE
## 1134          NA        NA          0   FALSE
## 1135          NA        NA          0   FALSE
## 1136          NA        NA          0   FALSE
## 1137          NA        NA          0   FALSE
## 1474          NA        NA          0   FALSE
## 1475          NA        NA          0   FALSE
## 1476          NA        NA          0   FALSE
## 1477          NA        NA          0   FALSE
## 2038          NA        NA          0   FALSE
## 2039          NA        NA          0   FALSE
## 2040          NA        NA          0   FALSE
## 2754          NA        NA          0   FALSE
```

En este caso se opta por eliminar los registros del conjunto de datos.

```
datos <- na.omit(datos)
colSums(is.na(datos))
```

```
##          Administrative Administrative_Duration           Informational
##                       0                       0                       0
##  Informational_Duration          ProductRelated ProductRelated_Duration
##                       0                       0                       0
##             BounceRates               ExitRates              PageValues
##                       0                       0                       0
##                 Revenue
##                       0
```

Se obtienen algunas estadísticas básicas:

```
summary(datos)
```

```
##  Administrative   Administrative_Duration Informational
##  Min.   : 0.000   Min.   :  -1.00         Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.:   0.00         1st Qu.: 0.000
##  Median : 1.000   Median :   8.00         Median : 0.000
##  Mean   : 2.318   Mean   :  80.91         Mean   : 0.504
##  3rd Qu.: 4.000   3rd Qu.:  93.50         3rd Qu.: 0.000
##  Max.   :27.000   Max.   :3398.75         Max.   :24.000
##  Informational_Duration ProductRelated   ProductRelated_Duration
##  Min.   :  -1.00        Min.   :  0.00   Min.   :   -1.0
##  1st Qu.:   0.00        1st Qu.:  7.00   1st Qu.:  185.0
##  Median :   0.00        Median : 18.00   Median :  599.8
##  Mean   :  34.51        Mean   : 31.76   Mean   : 1196.0
```

```
##  3rd Qu.:   0.00        3rd Qu.: 38.00   3rd Qu.: 1466.5
##  Max.   :2549.38        Max.   :705.00   Max.   :63973.5
##   BounceRates         ExitRates         PageValues       Revenue
##  Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   FALSE:10408
##  1st Qu.:0.000000   1st Qu.:0.01429   1st Qu.:  0.000   TRUE : 1908
##  Median :0.003119   Median :0.02512   Median :  0.000
##  Mean   :0.022152   Mean   :0.04300   Mean   :  5.896
##  3rd Qu.:0.016684   3rd Qu.:0.05000   3rd Qu.:  0.000
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764
```

Los valores negativos en algunos atributos de duración tienen valore negativos, lo que no tiene sentido para valores de tiempo, por lo que podría tratarse de un valor que indica que el dato no se registró. Algunos ejemplos de registros son:

```
datos[datos$Administrative_Duration <0,]
```

```
##      Administrative Administrative_Duration Informational
## 3                 0                     -1             0
## 7                 0                     -1             0
## 8                 1                     -1             0
## 17                0                     -1             0
## 22                0                     -1             0
## 25                0                     -1             0
## 50                0                     -1             0
## 51                0                     -1             0
## 65                0                     -1             0
## 133               0                     -1             0
## 141               0                     -1             0
## 182               0                     -1             0
## 183               0                     -1             0
## 253               0                     -1             0
## 384               0                     -1             0
## 533               0                     -1             0
## 541               0                     -1             0
## 563               0                     -1             0
## 592               0                     -1             0
## 639               0                     -1             0
## 2047              0                     -1             0
## 2053              0                     -1             0
## 2062              0                     -1             0
## 4800              0                     -1             0
## 4916              0                     -1             0
## 5096              0                     -1             0
## 5097              0                     -1             0
## 5108              0                     -1             0
## 5125              1                     -1             0
## 6261              0                     -1             0
## 7211              0                     -1             0
## 8053              0                     -1             0
## 8637              0                     -1             0
##      Informational_Duration ProductRelated ProductRelated_Duration
## 3                        -1              1                      -1
## 7                        -1              1                      -1
## 8                        -1              1                      -1
## 17                       -1              1                      -1
```

```
## 22                      -1              1                      -1
## 25                      -1              1                      -1
## 50                      -1              1                      -1
## 51                      -1              1                      -1
## 65                      -1              1                      -1
## 133                     -1              1                      -1
## 141                     -1              1                      -1
## 182                     -1              1                      -1
## 183                     -1              1                      -1
## 253                     -1              1                      -1
## 384                     -1              1                      -1
## 533                     -1              1                      -1
## 541                     -1              1                      -1
## 563                     -1              1                      -1
## 592                     -1              1                      -1
## 639                     -1              1                      -1
## 2047                    -1              1                      -1
## 2053                    -1              1                      -1
## 2062                    -1              1                      -1
## 4800                    -1              1                      -1
## 4916                    -1              1                      -1
## 5096                    -1              1                      -1
## 5097                    -1              1                      -1
## 5108                    -1              1                      -1
## 5125                    -1              1                      -1
## 6261                    -1              1                      -1
## 7211                    -1              1                      -1
## 8053                    -1              1                      -1
## 8637                    -1              1                      -1
##      BounceRates  ExitRates PageValues Revenue
## 3            0.2 0.20000000          0   FALSE
## 7            0.2 0.20000000          0   FALSE
## 8            0.2 0.20000000          0   FALSE
## 17           0.2 0.20000000          0   FALSE
## 22           0.2 0.20000000          0   FALSE
## 25           0.2 0.20000000          0   FALSE
## 50           0.2 0.20000000          0   FALSE
## 51           0.2 0.20000000          0   FALSE
## 65           0.2 0.20000000          0   FALSE
## 133          0.2 0.20000000          0   FALSE
## 141          0.2 0.20000000          0   FALSE
## 182          0.2 0.20000000          0   FALSE
## 183          0.2 0.20000000          0   FALSE
## 253          0.2 0.20000000          0   FALSE
## 384          0.2 0.20000000          0   FALSE
## 533          0.2 0.20000000          0   FALSE
## 541          0.2 0.20000000          0   FALSE
## 563          0.0 0.06666667          0   FALSE
## 592          0.2 0.20000000          0   FALSE
## 639          0.2 0.20000000          0   FALSE
## 2047         0.0 0.10000000          0   FALSE
## 2053         0.2 0.20000000          0   FALSE
## 2062         0.2 0.20000000          0   FALSE
## 4800         0.2 0.20000000          0   FALSE
```

```
## 4916         0.2 0.20000000          0   FALSE
## 5096         0.2 0.20000000          0   FALSE
## 5097         0.2 0.20000000          0   FALSE
## 5108         0.2 0.20000000          0   FALSE
## 5125         0.0 0.06666667          0   FALSE
## 6261         0.2 0.20000000          0   FALSE
## 7211         0.2 0.20000000          0   FALSE
## 8053         0.2 0.20000000          0   FALSE
## 8637         0.2 0.20000000          0   FALSE
```

Como se puede ver, en muschos de los casos, se tiene un valor 0 para el tipo de pagina y un valor de -1 para la duración. Por lo tanto, se reemplazarán los valores -1 por 0 cunado el tipo de página sea cero:

```
datos$Administrative_Duration  <- ifelse(datos$Administrative == 0, 0,datos$Administrative_Duration)
datos$Informational_Duration <- ifelse(datos$Informational == 0, 0,datos$Informational_Duration)
datos$ProductRelated_Duration <- ifelse(datos$ProductRelated == 0, 0,datos$ProductRelated_Duration)

summary(datos)
```

```
##  Administrative   Administrative_Duration Informational
##  Min.   : 0.000   Min.   :  -1.00         Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.:   0.00         1st Qu.: 0.000
##  Median : 1.000   Median :   8.00         Median : 0.000
##  Mean   : 2.318   Mean   :  80.91         Mean   : 0.504
##  3rd Qu.: 4.000   3rd Qu.:  93.50         3rd Qu.: 0.000
##  Max.   :27.000   Max.   :3398.75         Max.   :24.000
##  Informational_Duration ProductRelated   ProductRelated_Duration
##  Min.   :   0.00        Min.   :  0.00   Min.   :   -1.0
##  1st Qu.:   0.00        1st Qu.:  7.00   1st Qu.:  185.0
##  Median :   0.00        Median : 18.00   Median :  599.8
##  Mean   :  34.51        Mean   : 31.76   Mean   : 1196.0
##  3rd Qu.:   0.00        3rd Qu.: 38.00   3rd Qu.: 1466.5
##  Max.   :2549.38        Max.   :705.00   Max.   :63973.5
##   BounceRates         ExitRates         PageValues       Revenue
##  Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   FALSE:10408
##  1st Qu.:0.000000   1st Qu.:0.01429   1st Qu.:  0.000   TRUE : 1908
##  Median :0.003119   Median :0.02512   Median :  0.000
##  Mean   :0.022152   Mean   :0.04300   Mean   :  5.896
##  3rd Qu.:0.016684   3rd Qu.:0.05000   3rd Qu.:  0.000
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764
```

Existen todavia casos de valores negativos de duración:

```
datos[datos$Administrative_Duration <0,]
```

```
##      Administrative Administrative_Duration Informational
## 8                 1                     -1             0
## 5125              1                     -1             0
##      Informational_Duration ProductRelated ProductRelated_Duration
## 8                         0              1                      -1
## 5125                      0              1                      -1
##      BounceRates  ExitRates PageValues Revenue
## 8            0.2 0.20000000          0   FALSE
## 5125         0.0 0.06666667          0   FALSE
```

```
datos[datos$ProductRelated_Duration < 0,]
```

```
##      Administrative Administrative_Duration Informational
## 3                 0                       0             0
## 7                 0                       0             0
## 8                 1                      -1             0
## 17                0                       0             0
## 22                0                       0             0
## 25                0                       0             0
## 50                0                       0             0
## 51                0                       0             0
## 65                0                       0             0
## 133               0                       0             0
## 141               0                       0             0
## 182               0                       0             0
## 183               0                       0             0
## 253               0                       0             0
## 384               0                       0             0
## 533               0                       0             0
## 541               0                       0             0
## 563               0                       0             0
## 592               0                       0             0
## 639               0                       0             0
## 2047              0                       0             0
## 2053              0                       0             0
## 2062              0                       0             0
## 4800              0                       0             0
## 4916              0                       0             0
## 5096              0                       0             0
## 5097              0                       0             0
## 5108              0                       0             0
## 5125              1                      -1             0
## 6261              0                       0             0
## 7211              0                       0             0
## 8053              0                       0             0
## 8637              0                       0             0
##      Informational_Duration ProductRelated ProductRelated_Duration
## 3                         0              1                      -1
## 7                         0              1                      -1
## 8                         0              1                      -1
## 17                        0              1                      -1
## 22                        0              1                      -1
## 25                        0              1                      -1
## 50                        0              1                      -1
## 51                        0              1                      -1
## 65                        0              1                      -1
## 133                       0              1                      -1
## 141                       0              1                      -1
## 182                       0              1                      -1
## 183                       0              1                      -1
## 253                       0              1                      -1
## 384                       0              1                      -1
## 533                       0              1                      -1
## 541                       0              1                      -1
## 563                       0              1                      -1
## 592                       0              1                      -1
```

```
## 639                       0              1                       -1
## 2047                      0              1                       -1
## 2053                      0              1                       -1
## 2062                      0              1                       -1
## 4800                      0              1                       -1
## 4916                      0              1                       -1
## 5096                      0              1                       -1
## 5097                      0              1                       -1
## 5108                      0              1                       -1
## 5125                      0              1                       -1
## 6261                      0              1                       -1
## 7211                      0              1                       -1
## 8053                      0              1                       -1
## 8637                      0              1                       -1
##      BounceRates  ExitRates PageValues Revenue
## 3            0.2 0.20000000          0   FALSE
## 7            0.2 0.20000000          0   FALSE
## 8            0.2 0.20000000          0   FALSE
## 17           0.2 0.20000000          0   FALSE
## 22           0.2 0.20000000          0   FALSE
## 25           0.2 0.20000000          0   FALSE
## 50           0.2 0.20000000          0   FALSE
## 51           0.2 0.20000000          0   FALSE
## 65           0.2 0.20000000          0   FALSE
## 133          0.2 0.20000000          0   FALSE
## 141          0.2 0.20000000          0   FALSE
## 182          0.2 0.20000000          0   FALSE
## 183          0.2 0.20000000          0   FALSE
## 253          0.2 0.20000000          0   FALSE
## 384          0.2 0.20000000          0   FALSE
## 533          0.2 0.20000000          0   FALSE
## 541          0.2 0.20000000          0   FALSE
## 563          0.0 0.06666667          0   FALSE
## 592          0.2 0.20000000          0   FALSE
## 639          0.2 0.20000000          0   FALSE
## 2047         0.0 0.10000000          0   FALSE
## 2053         0.2 0.20000000          0   FALSE
## 2062         0.2 0.20000000          0   FALSE
## 4800         0.2 0.20000000          0   FALSE
## 4916         0.2 0.20000000          0   FALSE
## 5096         0.2 0.20000000          0   FALSE
## 5097         0.2 0.20000000          0   FALSE
## 5108         0.2 0.20000000          0   FALSE
## 5125         0.0 0.06666667          0   FALSE
## 6261         0.2 0.20000000          0   FALSE
## 7211         0.2 0.20000000          0   FALSE
## 8053         0.2 0.20000000          0   FALSE
## 8637         0.2 0.20000000          0   FALSE
```

En estos casos se realizará una imputación de valores, tomando la duraciòn promedio de Administrative cuando toma el valor 1 y de manera similar para el caso de Related Product:

```
mediaAdministrativeDuration <- mean(datos$Administrative_Duration[datos$Administrative==1 & datos$Admini
mediaAdministrativeDuration
```

```
## [1] 47.07236
```

```
mediaProductRelatedDuration <- mean(datos$ProductRelated_Duration[datos$ProductRelated==1 & datos$Produ
mediaProductRelatedDuration
```

```
## [1] 152.4028
```

```
datos$Administrative_Duration[datos$Administrative_Duration<0] =mediaAdministrativeDuration
datos$ProductRelated_Duration[datos$ProductRelated_Duration<0] = mediaProductRelatedDuration
```

Se vuelve a examinar el conjnto de datos

```
summary(datos)
```

```
##  Administrative   Administrative_Duration Informational
##  Min.   : 0.000   Min.   :   0.00         Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.:   0.00         1st Qu.: 0.000
##  Median : 1.000   Median :   8.00         Median : 0.000
##  Mean   : 2.318   Mean   :  80.92         Mean   : 0.504
##  3rd Qu.: 4.000   3rd Qu.:  93.50         3rd Qu.: 0.000
##  Max.   :27.000   Max.   :3398.75         Max.   :24.000
##  Informational_Duration ProductRelated   ProductRelated_Duration
##  Min.   :   0.00        Min.   :  0.00   Min.   :    0.0
##  1st Qu.:   0.00        1st Qu.:  7.00   1st Qu.:  185.0
##  Median :   0.00        Median : 18.00   Median :  599.8
##  Mean   :  34.51        Mean   : 31.76   Mean   : 1196.5
##  3rd Qu.:   0.00        3rd Qu.: 38.00   3rd Qu.: 1466.5
##  Max.   :2549.38        Max.   :705.00   Max.   :63973.5
##   BounceRates          ExitRates         PageValues       Revenue
##  Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   FALSE:10408
##  1st Qu.:0.000000   1st Qu.:0.01429   1st Qu.:  0.000   TRUE : 1908
##  Median :0.003119   Median :0.02512   Median :  0.000
##  Mean   :0.022152   Mean   :0.04300   Mean   :  5.896
##  3rd Qu.:0.016684   3rd Qu.:0.05000   3rd Qu.:  0.000
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764
```

3.2 Identificación y tratamiento de valores extremos.

Obtenemos las estadísticas básicas de los atributos del conjunto de datos:

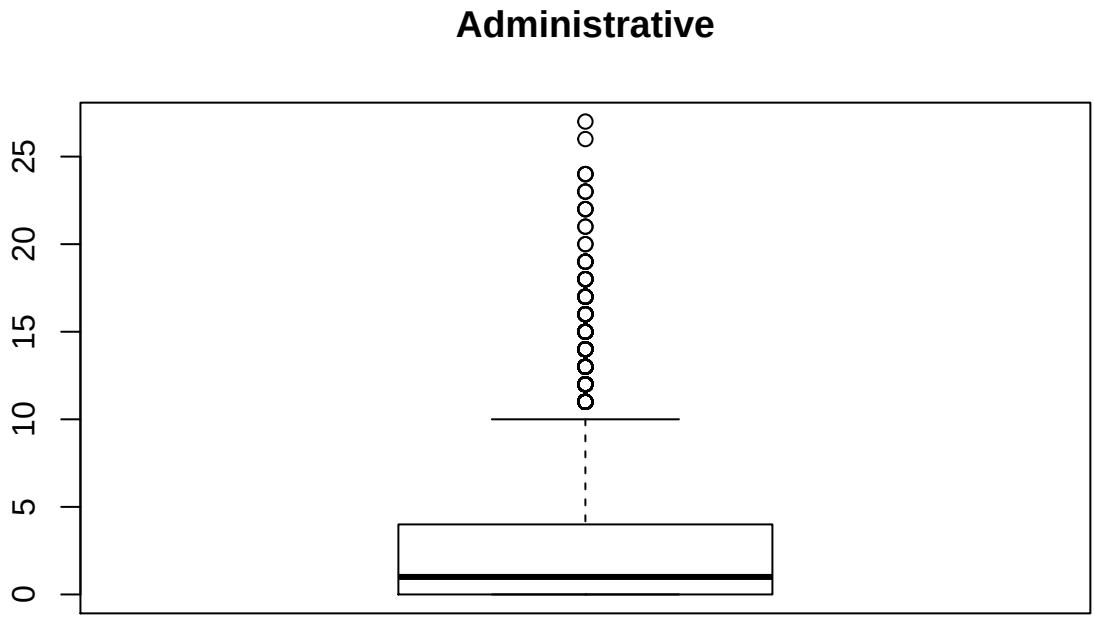
```
summary(datos)
```

```
##  Administrative   Administrative_Duration Informational
##  Min.   : 0.000   Min.   :   0.00         Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.:   0.00         1st Qu.: 0.000
##  Median : 1.000   Median :   8.00         Median : 0.000
##  Mean   : 2.318   Mean   :  80.92         Mean   : 0.504
##  3rd Qu.: 4.000   3rd Qu.:  93.50         3rd Qu.: 0.000
##  Max.   :27.000   Max.   :3398.75         Max.   :24.000
##  Informational_Duration ProductRelated   ProductRelated_Duration
##  Min.   :   0.00        Min.   :  0.00   Min.   :    0.0
##  1st Qu.:   0.00        1st Qu.:  7.00   1st Qu.:  185.0
##  Median :   0.00        Median : 18.00   Median :  599.8
##  Mean   :  34.51        Mean   : 31.76   Mean   : 1196.5
##  3rd Qu.:   0.00        3rd Qu.: 38.00   3rd Qu.: 1466.5
##  Max.   :2549.38        Max.   :705.00   Max.   :63973.5
```

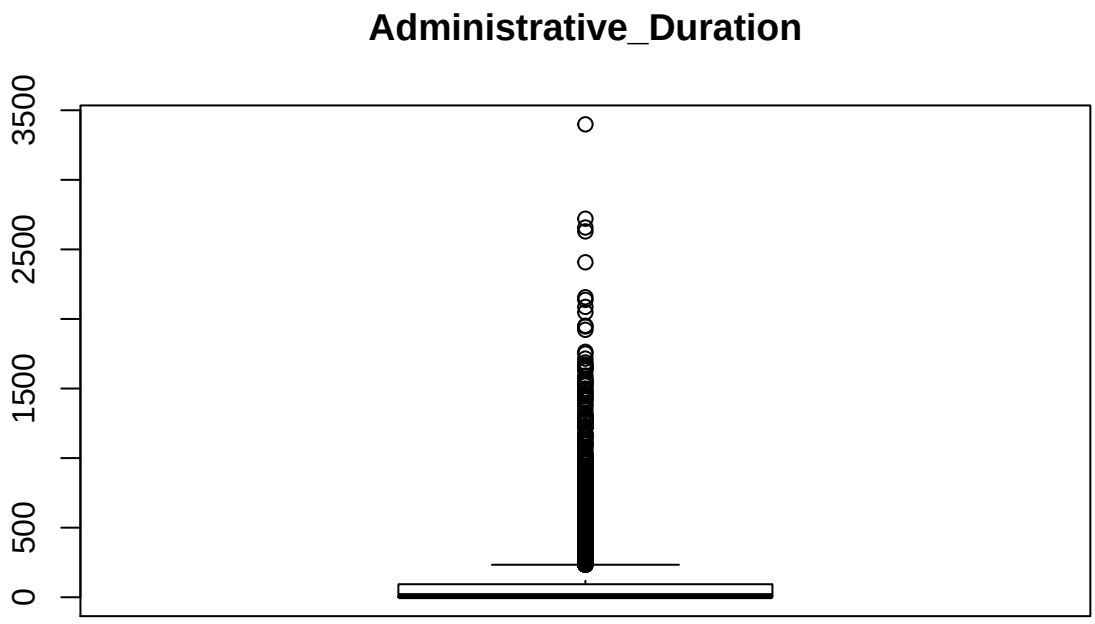
```
##   BounceRates         ExitRates         PageValues       Revenue
##  Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   FALSE:10408
##  1st Qu.:0.000000   1st Qu.:0.01429   1st Qu.:  0.000   TRUE : 1908
##  Median :0.003119   Median :0.02512   Median :  0.000
##  Mean   :0.022152   Mean   :0.04300   Mean   :  5.896
##  3rd Qu.:0.016684   3rd Qu.:0.05000   3rd Qu.:  0.000
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764
```

Mediante diagramas de caja y la librería outliers, se pueden analizar los valores que toman las diferentes variables numéricas.

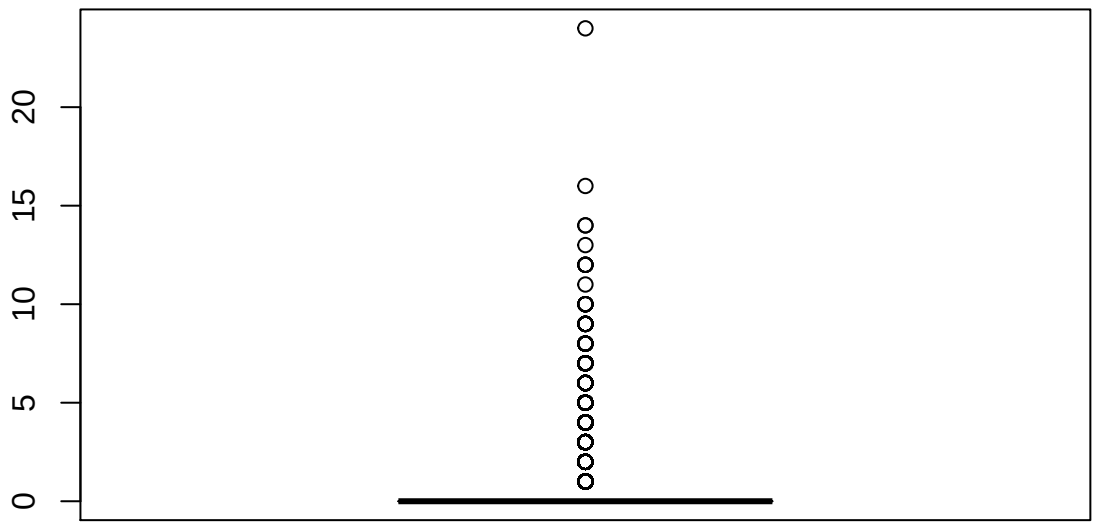
```
library(outliers)
for(atributo in atribNumericos){
  boxplot(datos[,atributo], main=atributo)
  print(outliers::outlier(x = datos[atributo]))
}
```



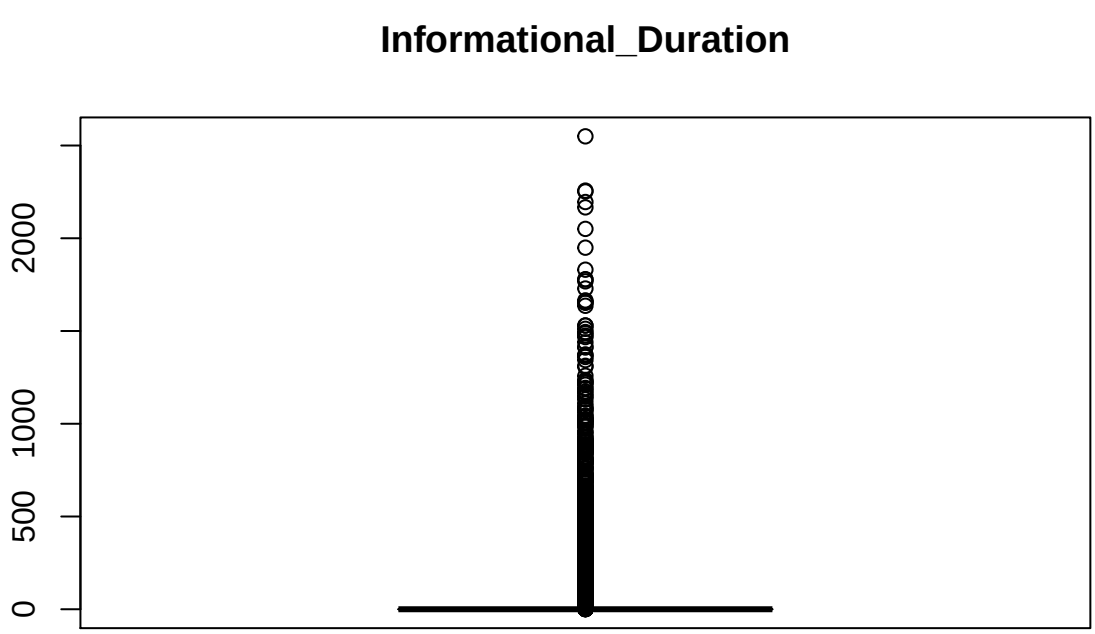

```
## Administrative
##             27
```



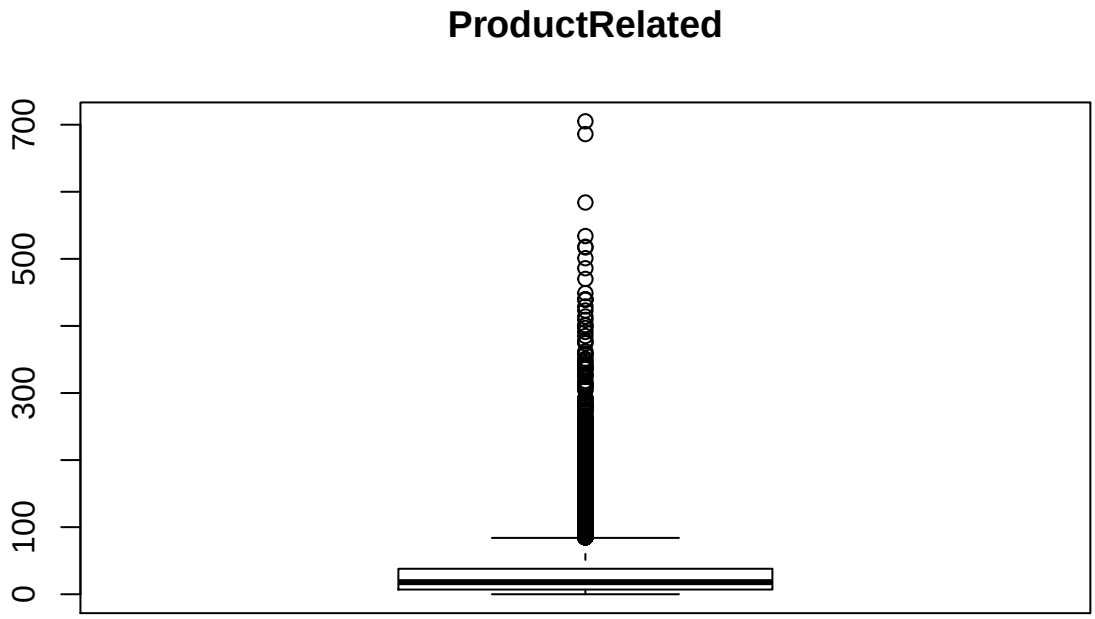

```
## Administrative_Duration
##                 3398.75
```



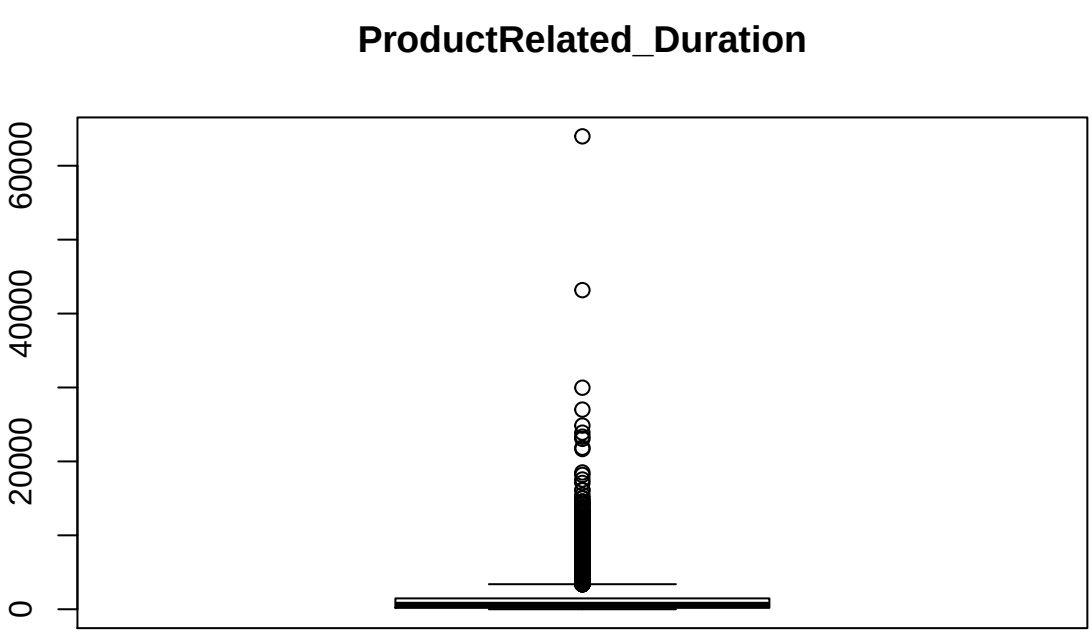

```
## Informational
##            24
```



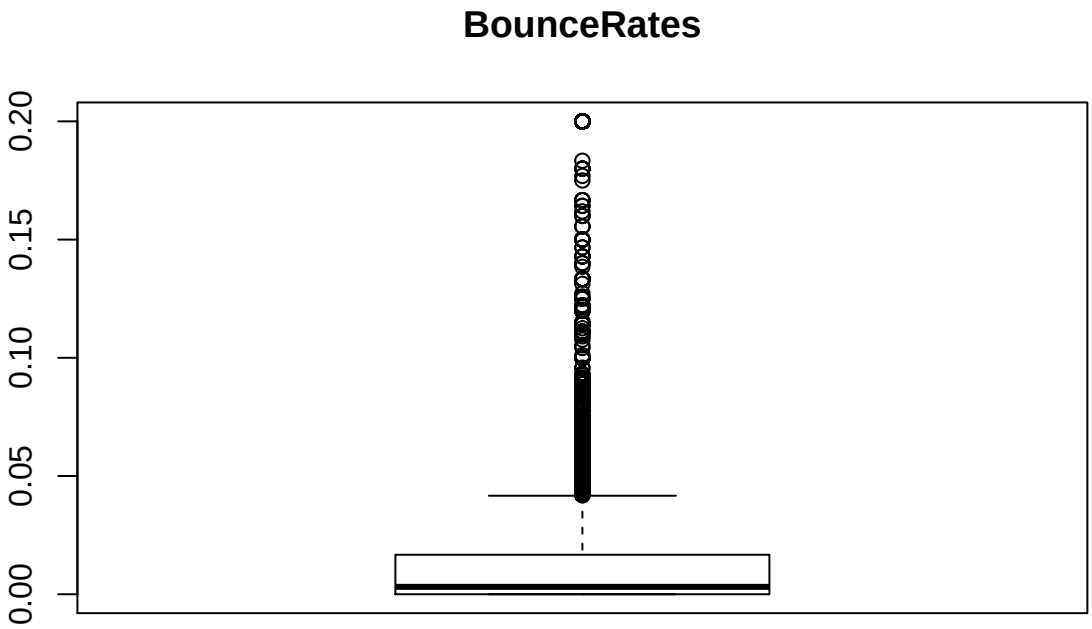

```
## Informational_Duration
##              2549.375
```



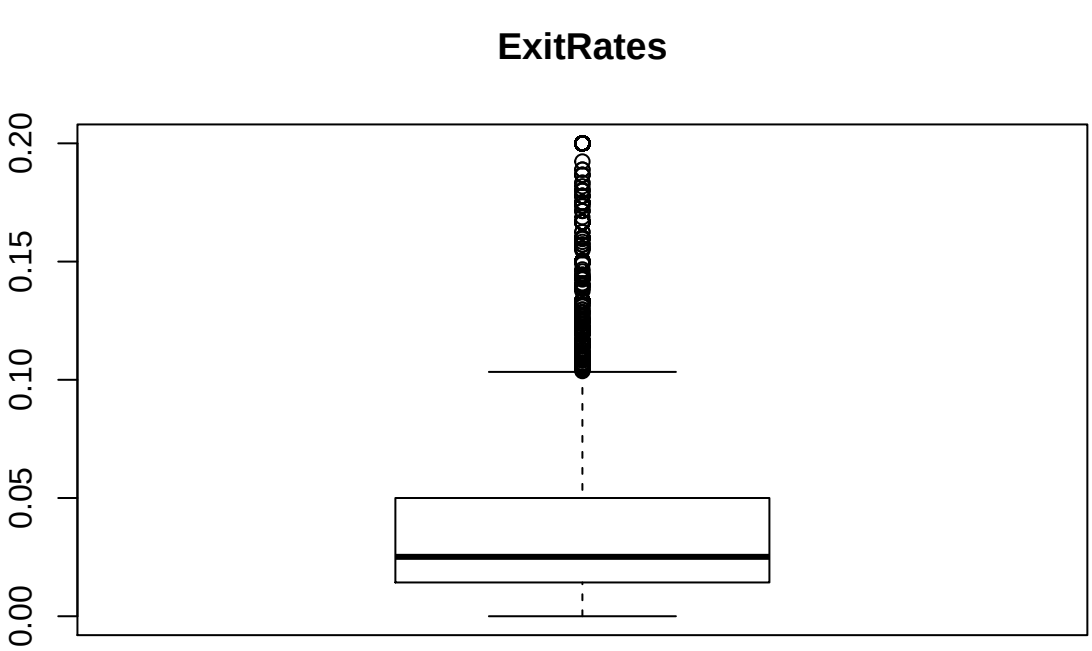

```
## ProductRelated
##            705
```



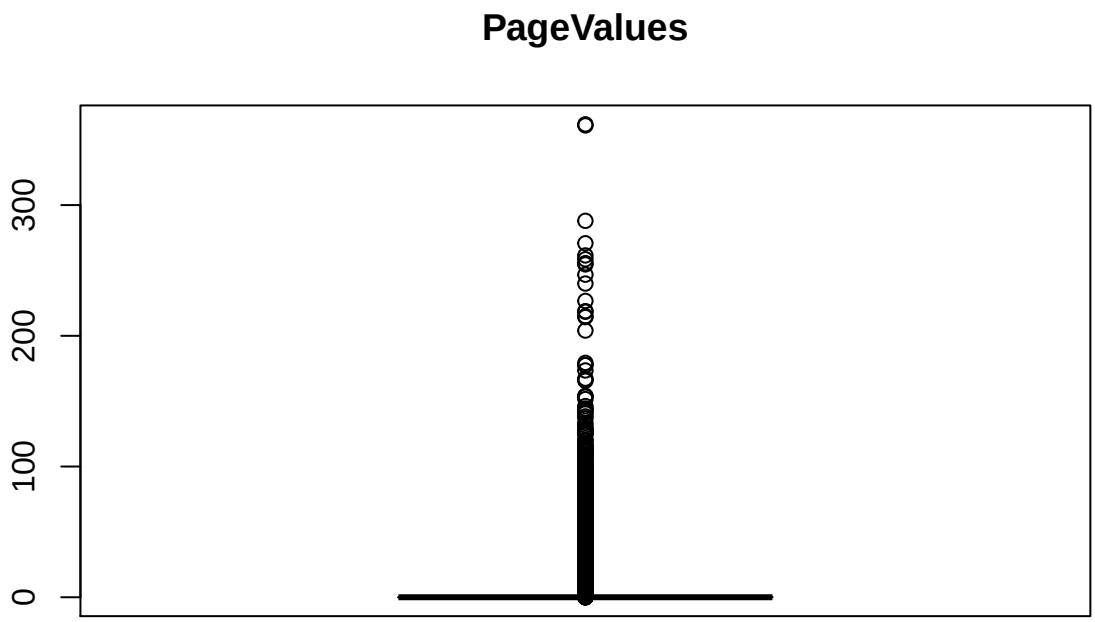

```
## ProductRelated_Duration
##                63973.52
```




```
## BounceRates
##         0.2
```




```
## ExitRates
##       0.2
```




```
## PageValues
##   361.7637
```

Como se puede ver, en el caso de la variable *Administrative_Duration*, se tiene un valor extremo de 3398.75:

```
datos[datos$Administrative_Duration >= 3000,]
```

```
##      Administrative Administrative_Duration Informational
## 8072              5                 3398.75             6
##      Informational_Duration ProductRelated ProductRelated_Duration
## 8072               2549.375            449                63973.52
##      BounceRates  ExitRates PageValues Revenue
## 8072 0.000764406 0.02770134          0   FALSE
```

Dado que este valor resulta anormal para el número de páginas administrativas (5) se eliminará el registro, esta eliminaciòn también elimina un valor extremo en la variable *Informational_Duration*

Otros valores muy notorios se dan en la varible *ProductRelated_duration", con una duración superior a los 40000

```
datos[datos$ProductRelated_Duration > 30000,]
```

```
##      Administrative Administrative_Duration Informational
## 5153             17                2629.254            24
## 8072              5                3398.750             6
##      Informational_Duration ProductRelated ProductRelated_Duration
## 5153               2050.433            705                43171.23
## 8072               2549.375            449                63973.52
##      BounceRates  ExitRates PageValues Revenue
```

```
## 5153 0.004851285 0.01543144   0.763829   FALSE
## 8072 0.000764406 0.02770134   0.000000   FALSE
```

Analizando los valores del resto de variables para las dos observaciones, se puede ver que estas dos transacciones ocurren en los meses de Mayo y Diciembre y que obedecen a una gran cantidad de páginas de productos relacionados visitadas, se podría explicar por ser meses relacionados fechas comerciales importantes, como lo son Navidad y día de la Madre. Analizando las observaciones de la variable *ProductRelated_Duration* por mes, se puede ver que incluso considerando que es el mes de diciembre, el valor de ProductRelated_Duration es muy alto incluso en función de las páginas visitadas:

Se decide eliminar el registro con duración de más de 63000

```
datos <- datos[datos$ProductRelated_Duration < 63000,]
```

Para analizar el efecto de esta eliminaciòn, se analiza de nuevo las estadísticas básicas del conjunto de datos

```
summary(datos)
```

```
##  Administrative   Administrative_Duration Informational
##  Min.   : 0.000   Min.   :   0.00         Min.   : 0.0000
##  1st Qu.: 0.000   1st Qu.:   0.00         1st Qu.: 0.0000
##  Median : 1.000   Median :   8.00         Median : 0.0000
##  Mean   : 2.318   Mean   :  80.65         Mean   : 0.5035
##  3rd Qu.: 4.000   3rd Qu.:  93.50         3rd Qu.: 0.0000
##  Max.   :27.000   Max.   :2720.50         Max.   :24.0000
##  Informational_Duration ProductRelated   ProductRelated_Duration
##  Min.   :   0.0         Min.   :  0.00   Min.   :    0.0
##  1st Qu.:   0.0         1st Qu.:  7.00   1st Qu.:  185.0
##  Median :   0.0         Median : 18.00   Median :  599.6
##  Mean   :  34.3         Mean   : 31.73   Mean   : 1191.3
##  3rd Qu.:   0.0         3rd Qu.: 38.00   3rd Qu.: 1465.8
##  Max.   :2256.9         Max.   :705.00   Max.   :43171.2
##   BounceRates         ExitRates         PageValues       Revenue
##  Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   FALSE:10407
##  1st Qu.:0.000000   1st Qu.:0.01429   1st Qu.:  0.000   TRUE : 1908
##  Median :0.003125   Median :0.02512   Median :  0.000
##  Mean   :0.022154   Mean   :0.04300   Mean   :  5.896
##  3rd Qu.:0.016701   3rd Qu.:0.05000   3rd Qu.:  0.000
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764
```

4 Análisis de los datos.

4.1 Grupos de datos

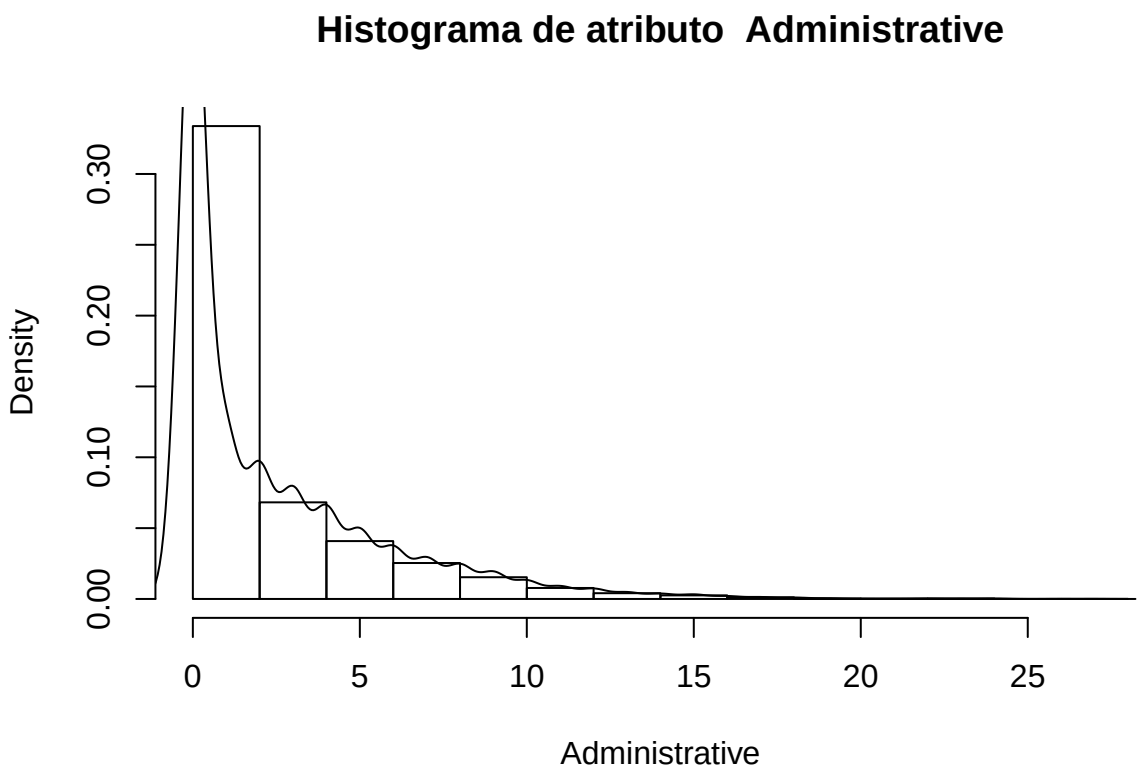
Par

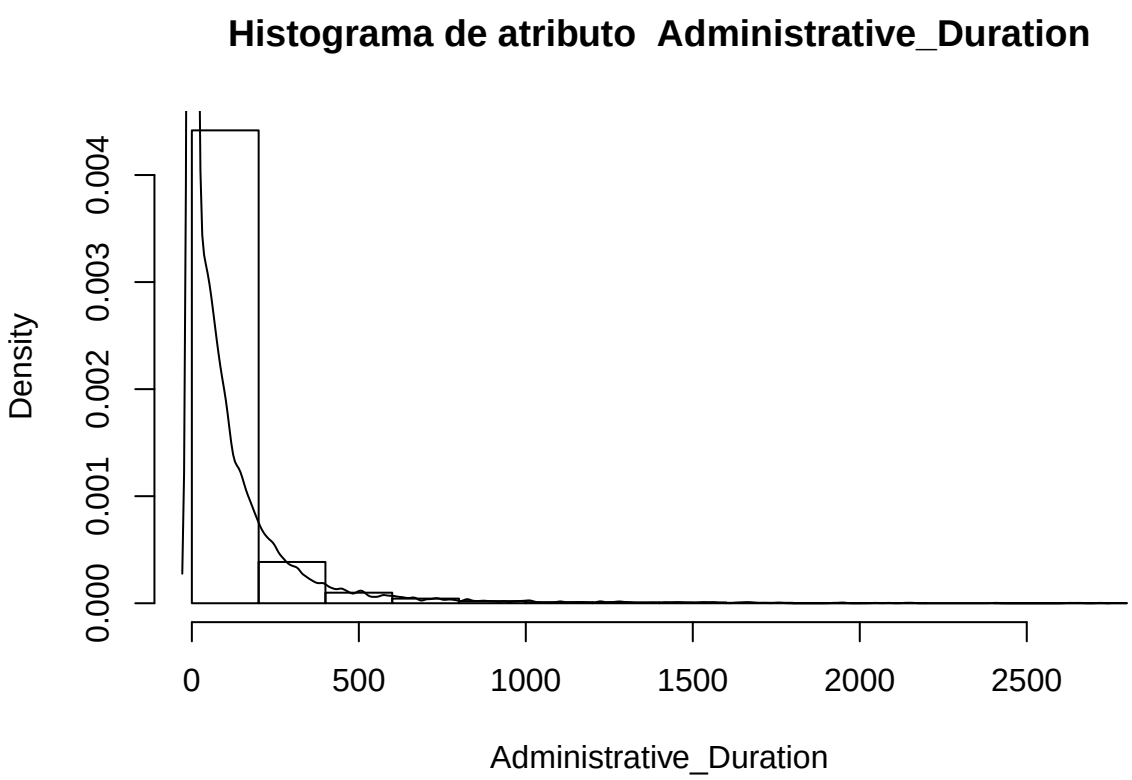
4.2 Comprobación de la normalidad y homogeneidad de la varianza.

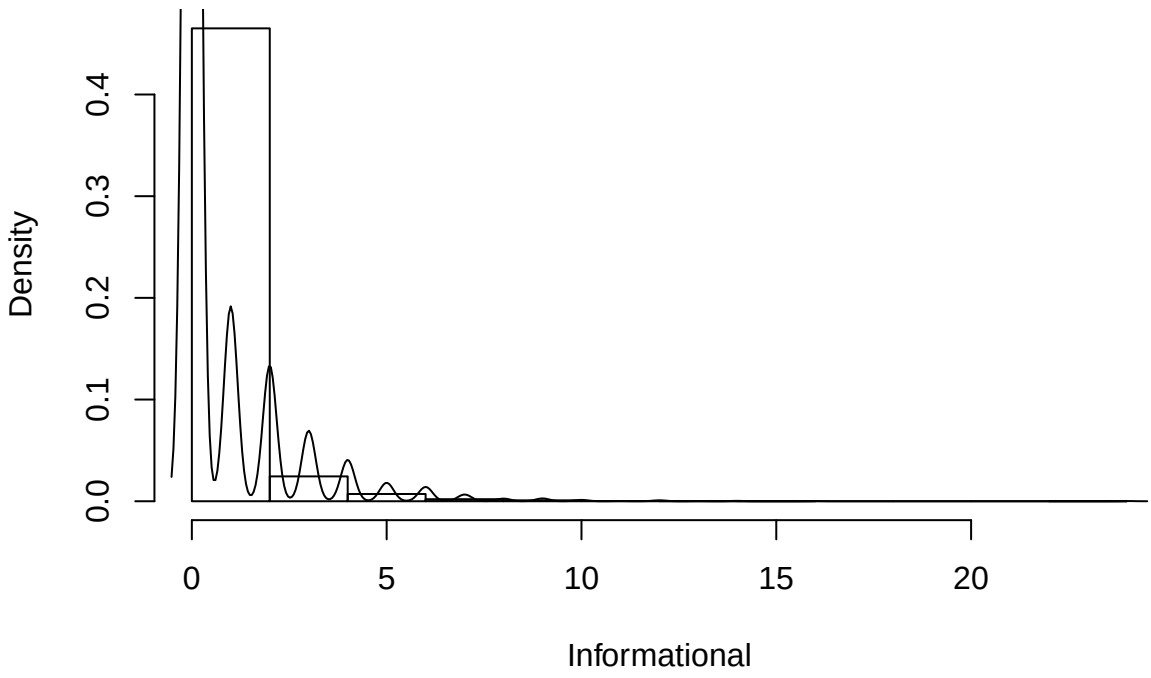
En primer lugar, se realiza una inspecciòn visual mediante la generación de histogramas para los diferentes atributos numéricos.

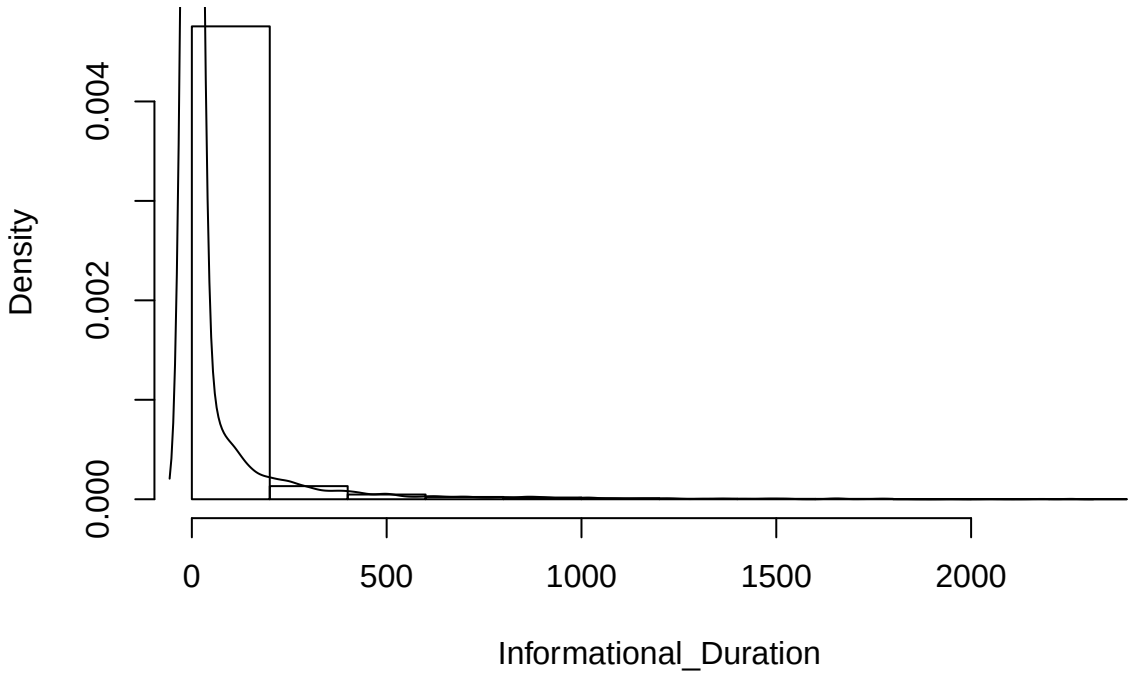
```
for(atributo in atribNumericos){
  hist(datos[,atributo],probability = T, main = paste("Histograma de atributo " , atributo), xlab = atri
```

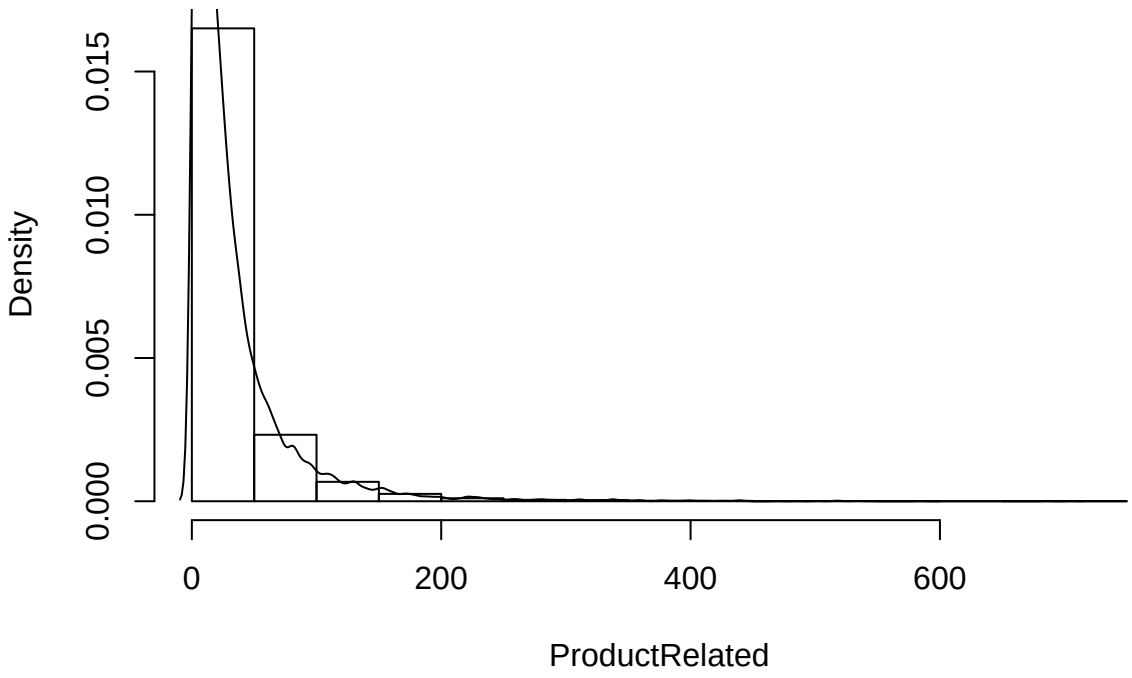
```
  lines(density(datos[,atributo]))
}
```

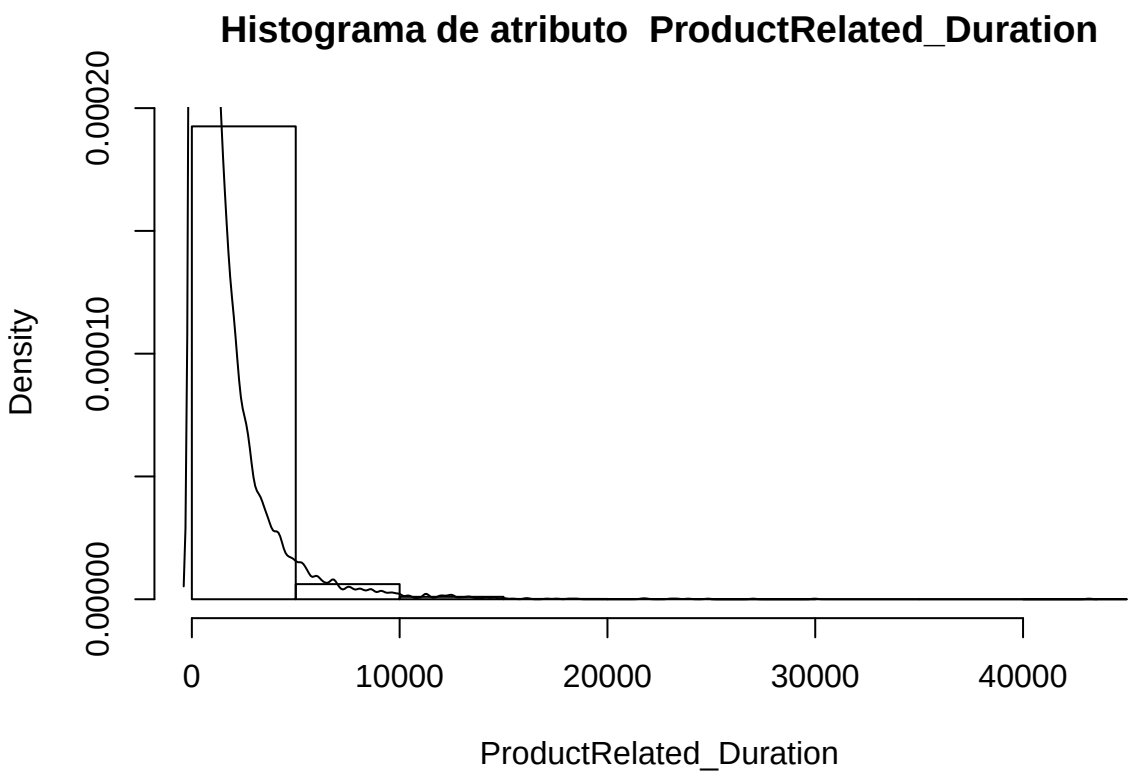


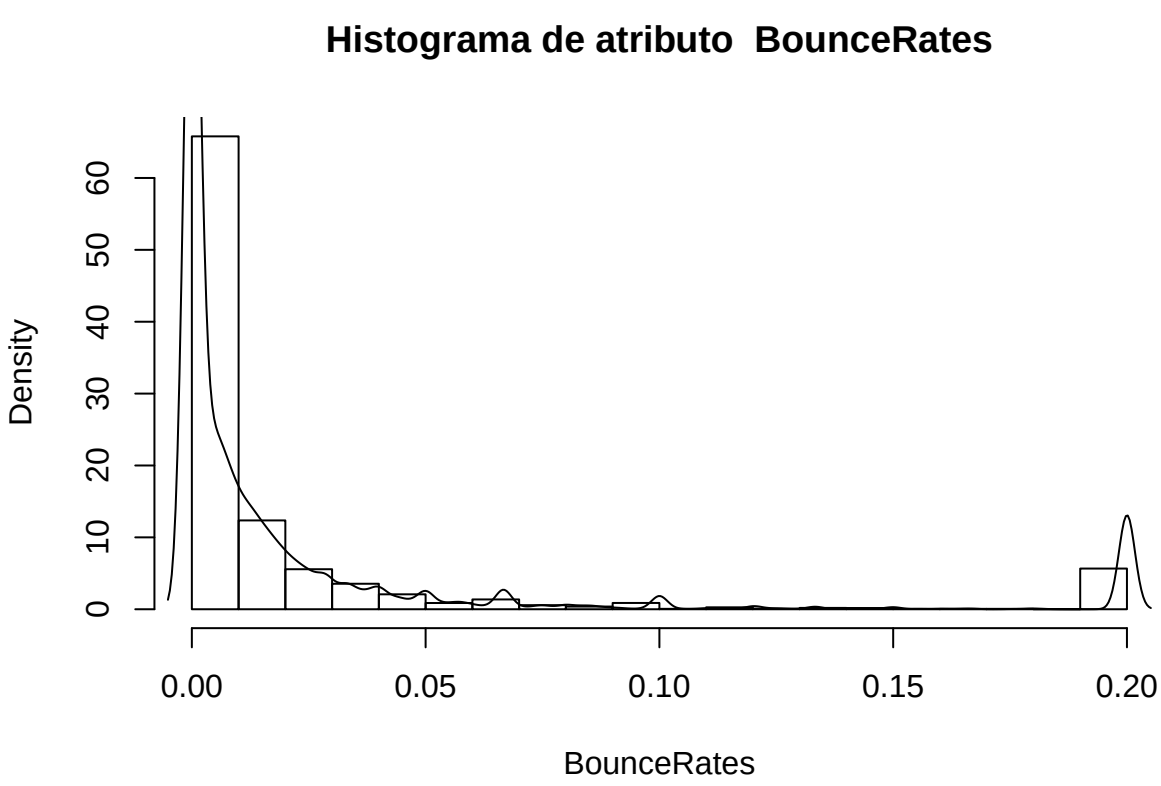

Histograma de atributo Administrative_Duration
Density
0.000
0.001
0.002
0.003
0.004
0
500
1000
1500
2000
2500
Administrative_Duration

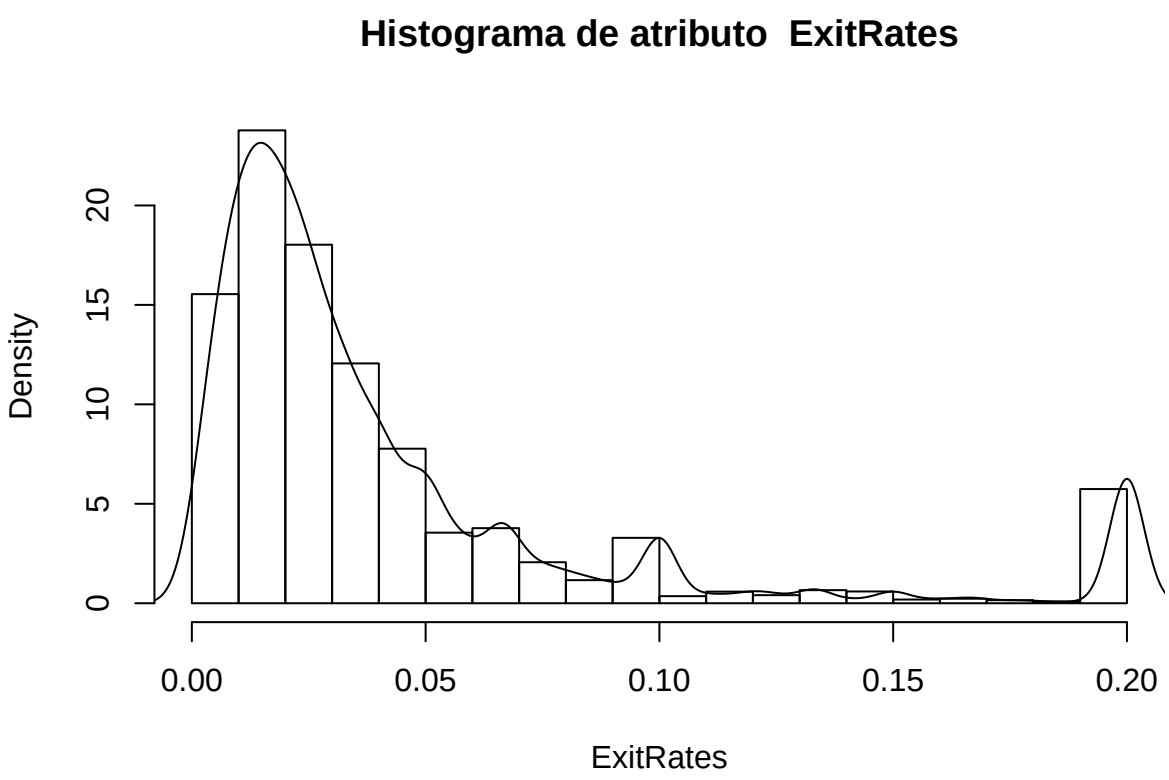

Histograma de atributo Informational
Density
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
Informational

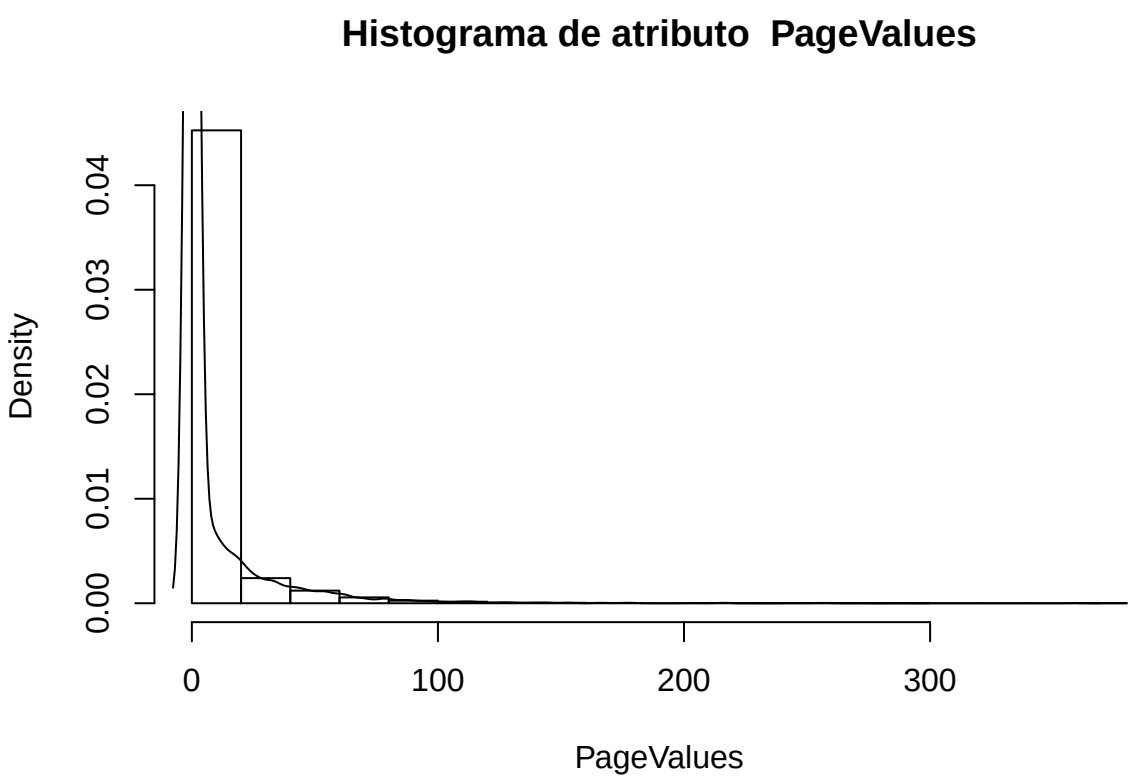

Histograma de atributo Informational_Duration
Density
0.000
0.002
0.004
0
500
1000
1500
2000
Informational_Duration


Histograma de atributo ProductRelated
Density
0.000
0.005
0.010
0.015
0
200
400
600
ProductRelated


Histograma de atributo ProductRelated_Duration
Density
0.00000
0.00010
0.00020
0
10000
20000
30000
40000
ProductRelated_Duration


Histograma de atributo BounceRates
Density
0
10
20
30
40
50
60
0.00
0.05
0.10
0.15
0.20
BounceRates


Histograma de atributo ExitRates
Density
0
5
10
15
20
0.00
0.05
0.10
0.15
0.20
ExitRates



La inspección visual sugiere que las variables no tienen una distribución normal, para comprobarlo se ejecuta una prueba Shapiro-Wilk para cada una de ellas. Se tomará una muestra de 5000 registros:

```
set.seed(202)
muestra <- dplyr::sample_n(datos,5000)
for(atributo in atribNumericos){
  valor_p = shapiro.test(muestra[,atributo])$p.value
  if(valor_p < 0.05){
    print(paste(atributo," no tiene una distribución normal, ","valor p:",valor_p))
  }
  else{
    print(paste(atributo," tiene una distribución normal, ","valor p:",valor_p))
  }
}
```

```
## [1] "Administrative  no tiene una distribución normal,  valor p: 1.23484411153285e-66"
## [1] "Administrative_Duration  no tiene una distribución normal,  valor p: 2.60720567633666e-80"
## [1] "Informational  no tiene una distribución normal,  valor p: 3.39003895636515e-82"
## [1] "Informational_Duration  no tiene una distribución normal,  valor p: 1.65502388058869e-88"
## [1] "ProductRelated  no tiene una distribución normal,  valor p: 4.1647438533107e-75"
## [1] "ProductRelated_Duration  no tiene una distribución normal,  valor p: 7.04415880538803e-77"
## [1] "BounceRates  no tiene una distribución normal,  valor p: 1.44791069205372e-80"
## [1] "ExitRates  no tiene una distribución normal,  valor p: 1.18187101699149e-69"
## [1] "PageValues  no tiene una distribución normal,  valor p: 3.72880291406678e-86"
```

4.3 Pruebas estadísticas

Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

4.3.1 Análisis de la correlación de las variables seleccionadas con el resultado

El siguiente análisis de correlación se realiza utilizando la prueba de Spearman ya que se ha determinado que los valores no siguen una distribución normal.

```
kruskal.test(data=datos, Informational~Revenue)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Informational by Revenue
## Kruskal-Wallis chi-squared = 159.65, df = 1, p-value < 2.2e-16
```

```
kruskal.test(data=datos, Informational_Duration~Revenue)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Informational_Duration by Revenue
## Kruskal-Wallis chi-squared = 154.8, df = 1, p-value < 2.2e-16
```

```
kruskal.test(data=datos, Administrative~Revenue)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Administrative by Revenue
## Kruskal-Wallis chi-squared = 345.96, df = 1, p-value < 2.2e-16
```

```
kruskal.test(data=datos, Administrative_Duration~Revenue)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Administrative_Duration by Revenue
## Kruskal-Wallis chi-squared = 329.6, df = 1, p-value < 2.2e-16
```

```
kruskal.test(data=datos, ProductRelated~Revenue)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  ProductRelated by Revenue
## Kruskal-Wallis chi-squared = 483.72, df = 1, p-value < 2.2e-16
```

```
kruskal.test(data=datos, ProductRelated_Duration~Revenue)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  ProductRelated_Duration by Revenue
## Kruskal-Wallis chi-squared = 575.66, df = 1, p-value < 2.2e-16
```

```
library("Hmisc")
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
res2 <- rcorr(as.matrix(datos[names(datos) %in% c(atribNumericos)]))
res2
```

```
##                          Administrative Administrative_Duration
## Administrative                     1.00                    0.61
## Administrative_Duration            0.61                    1.00
## Informational                      0.38                    0.30
## Informational_Duration             0.26                    0.22
## ProductRelated                     0.43                    0.28
## ProductRelated_Duration            0.39                    0.32
## BounceRates                       -0.22                   -0.15
## ExitRates                         -0.32                   -0.21
## PageValues                         0.10                    0.07
##                          Informational Informational_Duration
## Administrative                    0.38                   0.26
## Administrative_Duration           0.30                   0.22
## Informational                     1.00                   0.62
## Informational_Duration            0.62                   1.00
## ProductRelated                    0.37                   0.27
## ProductRelated_Duration           0.39                   0.32
## BounceRates                      -0.12                  -0.07
## ExitRates                        -0.16                  -0.11
## PageValues                        0.05                   0.03
##                          ProductRelated ProductRelated_Duration BounceRates
## Administrative                     0.43                    0.39       -0.22
## Administrative_Duration            0.28                    0.32       -0.15
## Informational                      0.37                    0.39       -0.12
## Informational_Duration             0.27                    0.32       -0.07
## ProductRelated                     1.00                    0.88       -0.20
## ProductRelated_Duration            0.88                    1.00       -0.19
## BounceRates                       -0.20                   -0.19        1.00
## ExitRates                         -0.29                   -0.26        0.91
## PageValues                         0.06                    0.06       -0.12
##                          ExitRates PageValues
## Administrative               -0.32       0.10
## Administrative_Duration      -0.21       0.07
## Informational                -0.16       0.05
## Informational_Duration       -0.11       0.03
## ProductRelated               -0.29       0.06
## ProductRelated_Duration      -0.26       0.06
```

```
## BounceRates                  0.91      -0.12
## ExitRates                    1.00      -0.17
## PageValues                  -0.17       1.00
##
## n= 12315
##
##
## P
##                          Administrative Administrative_Duration
## Administrative                           0e+00
## Administrative_Duration 0e+00
## Informational            0e+00           0e+00
## Informational_Duration   0e+00           0e+00
## ProductRelated           0e+00           0e+00
## ProductRelated_Duration  0e+00           0e+00
## BounceRates              0e+00           0e+00
## ExitRates                0e+00           0e+00
## PageValues               0e+00           0e+00
##                          Informational Informational_Duration
## Administrative           0e+00         0e+00
## Administrative_Duration  0e+00         0e+00
## Informational                          0e+00
## Informational_Duration   0e+00
## ProductRelated           0e+00         0e+00
## ProductRelated_Duration  0e+00         0e+00
## BounceRates              0e+00         0e+00
## ExitRates                0e+00         0e+00
## PageValues               0e+00         4e-04
##                          ProductRelated ProductRelated_Duration BounceRates
## Administrative           0e+00           0e+00                  0e+00
## Administrative_Duration  0e+00           0e+00                  0e+00
## Informational            0e+00           0e+00                  0e+00
## Informational_Duration   0e+00           0e+00                  0e+00
## ProductRelated                           0e+00                  0e+00
## ProductRelated_Duration  0e+00                                  0e+00
## BounceRates              0e+00           0e+00
## ExitRates                0e+00           0e+00                  0e+00
## PageValues               0e+00           0e+00                  0e+00
##                          ExitRates PageValues
## Administrative           0e+00     0e+00
## Administrative_Duration  0e+00     0e+00
## Informational            0e+00     0e+00
## Informational_Duration   0e+00     4e-04
## ProductRelated           0e+00     0e+00
## ProductRelated_Duration  0e+00     0e+00
## BounceRates              0e+00     0e+00
## ExitRates                          0e+00
## PageValues               0e+00
```

5 Predicción

Se plantea utilizar un modelo de regresión logística para predecir si un visitante realizará o no una compra, es decir si la variable *Revenue* tendrá un valor Verdadero.

Se divide el conjunto de datos en un conjunto de entrenamiento y uno de prueba:

```
library(caTools)
set.seed(123)
datosSelec = datos
split = sample.split(datosSelec$Revenue, SplitRatio = 0.75)
training_set = subset(datosSelec, split == TRUE)
test_set = subset(datosSelec, split == FALSE)
#La ultima columna es la variable dependiente
ultimaCol = dim(training_set)[2]
```

Seguidamente realizamos una normalizacion de los datos para las variables numéricas.

```
for(atributo in atribNumericos){
  training_set[,atributo] = scale(training_set[,atributo])
  test_set[,atributo] = scale(test_set[,atributo])
}
```

Creamos el clasificador

```
classifier = glm(formula = Revenue ~ .,
                 family = binomial,
                 data = training_set)
```

Se realiza la predicción

```
prob_pred = predict(classifier, type = 'response', newdata = test_set[-ultimaCol])
y_pred = ifelse(prob_pred > 0.5, T, F)
```

Se analiza el resultado en con una matriz de confusion

```
cm = table(test_set[, ultimaCol], y_pred)
cm
```

```
##        y_pred
##         FALSE TRUE
##   FALSE  2537   65
##   TRUE    283  194
```

6 Representación de los resultados a partir de tablas y gráficas.

7 Resolución del problema.

Con el modelo creado se puede realizar una predicción correcta del 88% (2731 de 3079 casos) de casos usando las variables seleccionadas.