UCI Shopping EDA

Aaron Tang

2026-02-09

Link to dataset: https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset

```
# Loading data
work_dir <- here()
raw_data <- read_csv(file.path(work_dir, "shopping", "data", "raw_data", "uci_shopping.csv"),
                     show_col_types =  FALSE)
head(raw_data)
```

```
## # A tibble: 6 x 18
##   Administrative Administrative_Duration Informational Informational_Duration
##            <dbl>                   <dbl>         <dbl>                   <dbl>
## 1              0                       0             0                       0
## 2              0                       0             0                       0
## 3              0                       0             0                       0
## 4              0                       0             0                       0
## 5              0                       0             0                       0
## 6              0                       0             0                       0
## # i 14 more variables: ProductRelated <dbl>, ProductRelated_Duration <dbl>,
## #   BounceRates <dbl>, ExitRates <dbl>, PageValues <dbl>, SpecialDay <dbl>,
## #   Month <chr>, OperatingSystems <dbl>, Browser <dbl>, Region <dbl>,
## #   TrafficType <dbl>, VisitorType <chr>, Weekend <lgl>, Revenue <lgl>
```

Description of Variables

Continuous

- `Administrative`: Number of pages visited about account management
- `Administrative Duration`: Total amount of time (in seconds) spent on account management related pages
- `Informational`: Number of pages visited about website, communication and address information of the shopping site
- `Informational Duration`: Total amount of time spent on informational pages
- `Product Related`: Number of pages visited about product related pages
- `Product Related Duration`: Total amount of time spent on product related pages

- `Bounce Rate`: Average bounce rate value of the pages visited

  - Formula: $BounceRate = 1 - EngagementRate$
  - $EngagementRate$ = percentage of sessions that either lasted longer than 10 seconds, had key events, or two or more screen or page views

  - Google Analytics assigns this value to every page of a website
  - This feature takes the average of all the bounce rates value from all the pages the user visits in a single session

- `Exit Rate`: Average exit rate value of the pages visited
  - Formula: $ExitRate = \frac{NumberOfExits}{NumberOfPageViews} \times 100$

- `Page Value`: Average page value of the pages visited
  - Formula: $PageValue = \frac{TotalRevenue + TotalGoalValue}{UniquePageViews}$
  - $TotalRevenue$ = amount of money generated from a website page (i.e. transaction page)
  - $TotalGoalValue$ = value assigned to a specific page that is defined by the business
  - $UniquePageViews$ = number of unique user visits, only counted once per session

- `Special Day`: Closeness of the site visiting time to a special day

Categorical

- `OperatingSystems`: Operating system of the visitor
- `Browser`: Browser of the visitor
- `Region`: Geographic region from which the session has been started by the visitor
- `TrafficType`: Traffic source by which the visitor has arrived at the wesbite
- `VisitorType`: Visitor type as "New Visitor", "Returning Visitor", and "Other"
- `Weekend`: Boolean value indicating whether the date of the visit is a weekend
- `Month`: Month value of the visit date
- `Revenue`: Class label indicating whether the visit has been finalized with a transaction

Summary

```
continuous_var <- c("Administrative", "Administrative_Duration", "Informational",
                    "Informational_Duration", "ProductRelated", "ProductRelated_Duration",
                    "BounceRates", "ExitRates", "PageValues")

categorical_var <- c("Month", "OperatingSystems", "Browser", "Region", "TrafficType",
                     "VisitorType", "Weekend", "Revenue")

shopping_uci <- raw_data %>%
  mutate(across(all_of(categorical_var), as.factor))

summary(shopping_uci)
```

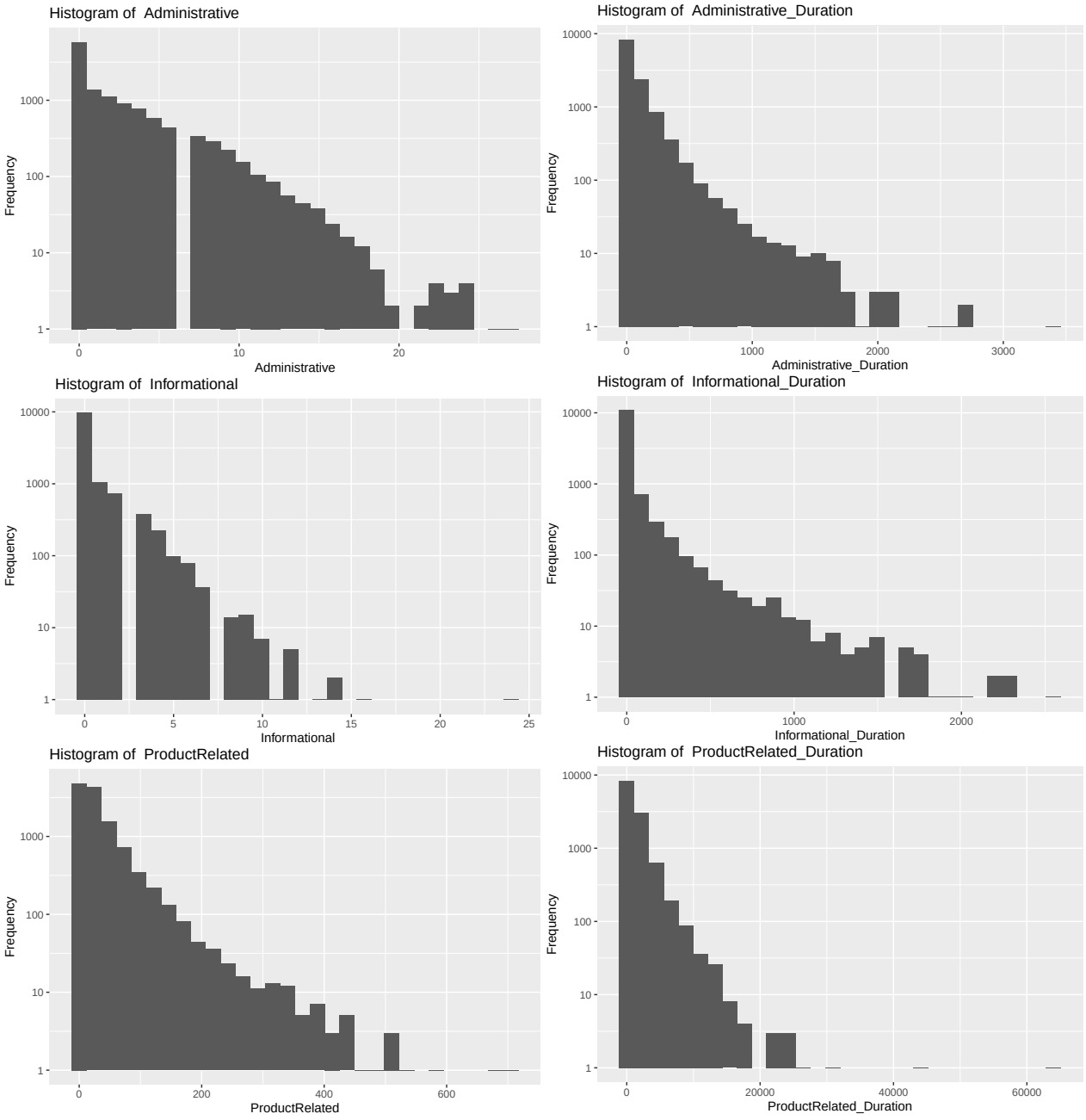
```
##  Administrative   Administrative_Duration Informational
##  Min.   : 0.000   Min.   :   0.00         Min.   : 0.0000
##  1st Qu.: 0.000   1st Qu.:   0.00         1st Qu.: 0.0000
##  Median : 1.000   Median :   7.50         Median : 0.0000
##  Mean   : 2.315   Mean   :  80.82         Mean   : 0.5036
```

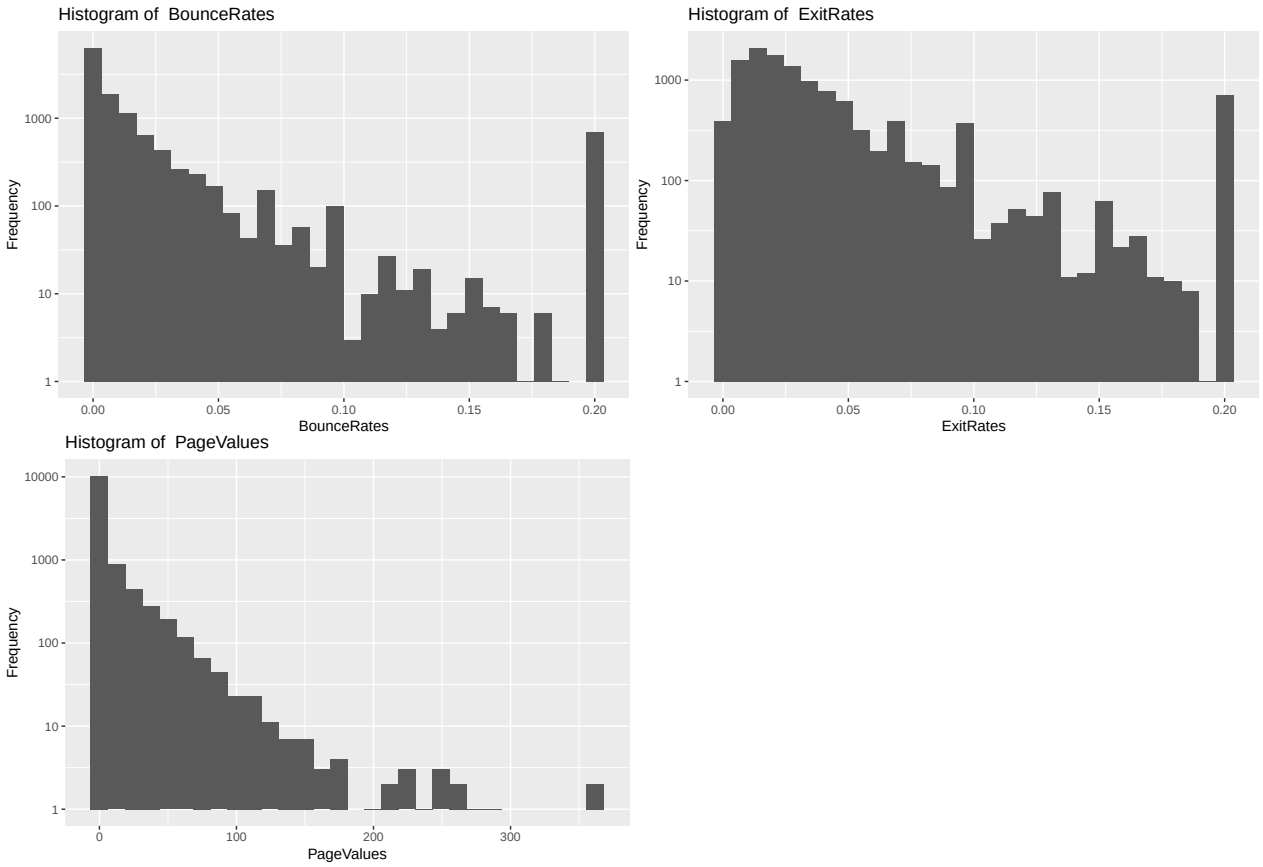
```
##  3rd Qu.: 4.000   3rd Qu.:  93.26         3rd Qu.: 0.0000
##  Max.   :27.000   Max.   :3398.75         Max.   :24.0000
##
##  Informational_Duration ProductRelated   ProductRelated_Duration
##  Min.   :   0.00        Min.   :  0.00   Min.   :    0.0
##  1st Qu.:   0.00        1st Qu.:  7.00   1st Qu.:  184.1
##  Median :   0.00        Median : 18.00   Median :  598.9
##  Mean   :  34.47        Mean   : 31.73   Mean   : 1194.8
##  3rd Qu.:   0.00        3rd Qu.: 38.00   3rd Qu.: 1464.2
##  Max.   :2549.38        Max.   :705.00   Max.   :63973.5
##
##   BounceRates         ExitRates         PageValues        SpecialDay
##  Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.01429   1st Qu.:  0.000   1st Qu.:0.00000
##  Median :0.003112   Median :0.02516   Median :  0.000   Median :0.00000
##  Mean   :0.022191   Mean   :0.04307   Mean   :  5.889   Mean   :0.06143
##  3rd Qu.:0.016813   3rd Qu.:0.05000   3rd Qu.:  0.000   3rd Qu.:0.00000
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764   Max.   :1.00000
##
##      Month      OperatingSystems    Browser         Region      TrafficType
##  May    :3364   2      :6601     2      :7961   1      :4780   2      :3913
##  Nov    :2998   1      :2585     1      :2462   3      :2403   1      :2451
##  Mar    :1907   3      :2555     4      : 736   4      :1182   3      :2052
##  Dec    :1727   4      : 478     5      : 467   2      :1136   4      :1069
##  Oct    : 549   8      :  79     6      : 174   6      : 805   13     : 738
##  Sep    : 448   6      :  19     10     : 163   7      : 761   10     : 450
##  (Other):1337   (Other):  13     (Other): 367   (Other):1263   (Other):1657
##            VisitorType     Weekend      Revenue
##  New_Visitor      : 1694   FALSE:9462   FALSE:10422
##  Other            :   85   TRUE :2868   TRUE : 1908
##  Returning_Visitor:10551
##
##
##
##
```

Bar Charts/Histograms of Variables

```r
# Histograms of Continuous Variables
par(mfrow = c(2, 3))

for (i in 1:length(continuous_var)) {
  current = shopping_uci[[continuous_var[i]]]
  plot <- ggplot(data = shopping_uci, mapping = aes(x = current)) +
    geom_histogram() +
    labs(
      y = "Frequency",
      x = continuous_var[i],
      title = paste("Histogram of ", continuous_var[i])
    ) +
    scale_y_log10()
```

```
  print(plot)
}
```

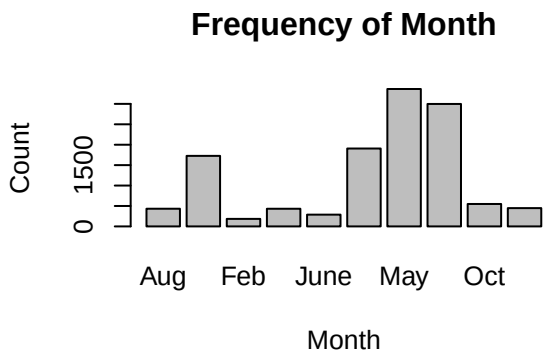

Histogram of Administrative
Frequency
Administrative
Histogram of Administrative_Duration
Frequency
Administrative_Duration
Histogram of Informational
Frequency
Informational
Histogram of Informational_Duration
Frequency
Informational_Duration
Histogram of ProductRelated
Frequency
ProductRelated
Histogram of ProductRelated_Duration
Frequency
ProductRelated_Duration

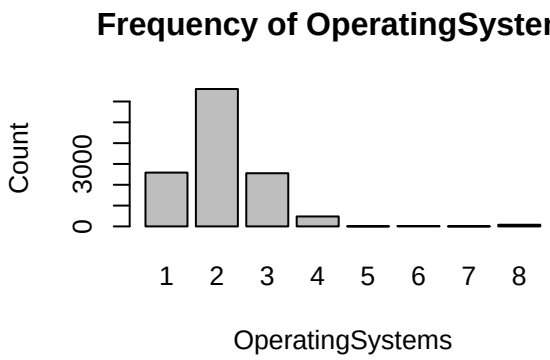


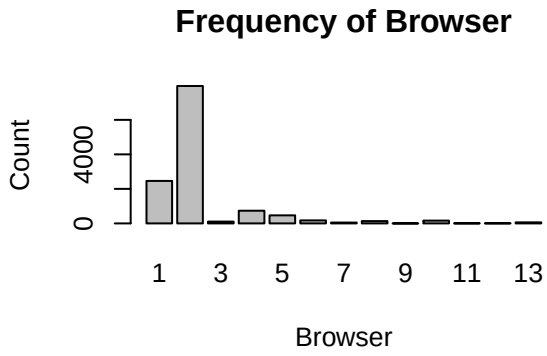
Notes:

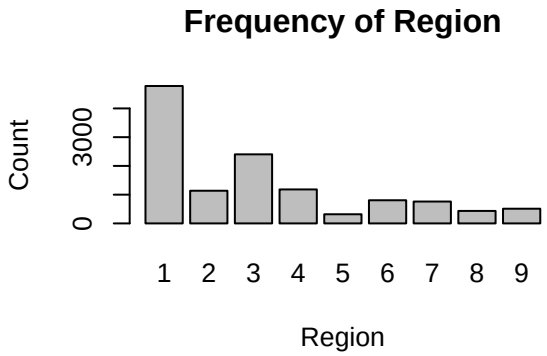
- I logged the scale of the y-axes because they were all heavily zero-inflated
- After log transforming the scale, it appears the duration graphs follow a somewhat exponential distribution
- The rates appear to be a little left skewed

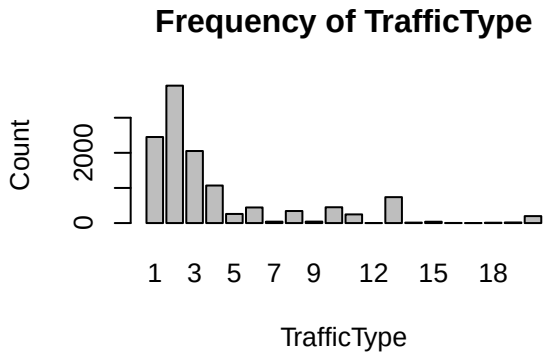
```
# Bar Charts of Categorical Variables
par(mfrow = c(2, 2))

for (i in 1:length(categorical_var)) {
  current_var_count <- shopping_uci %>%
    group_by(across(all_of(categorical_var[i]))) %>%
    summarize(count = n())

  barplot(height = current_var_count$count,
          names.arg = current_var_count %>% pull(1),
          xlab = categorical_var[i],
          ylab = "Count",
          main = paste("Frequency of", categorical_var[i]))
}
```

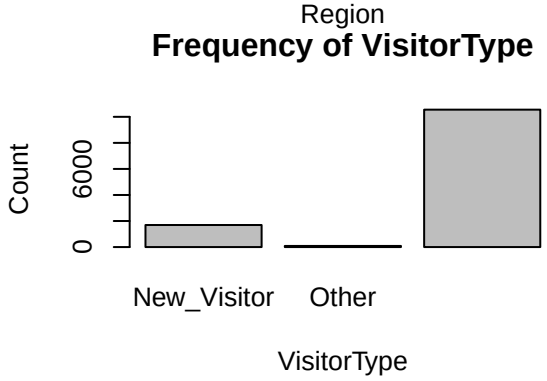

Frequency of Month
Count
0
1500
Aug
Feb
June
May
Oct
Month

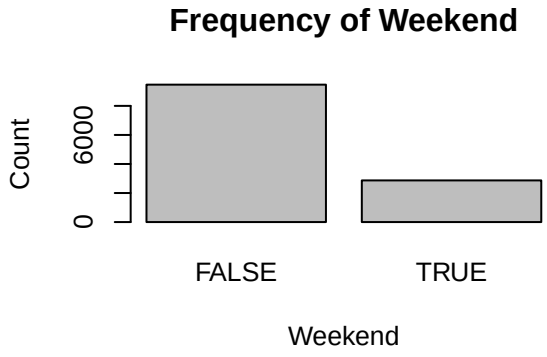

Frequency of OperatingSystems
Count
0
3000
1
2
3
4
5
6
7
8
OperatingSystems

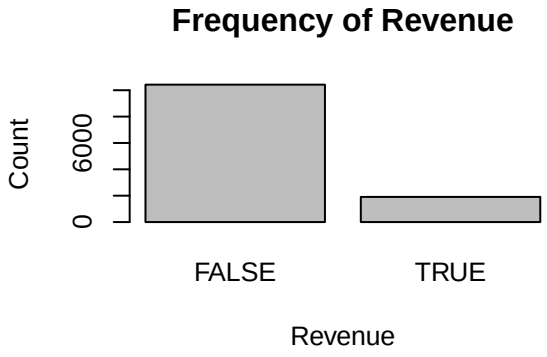

Frequency of Browser
Count
0
4000
1
3
5
7
9
11
13
Browser


Frequency of Region
Count
0
3000
1
2
3
4
5
6
7
8
9
Region


Frequency of TrafficType
Count
0
2000
1
3
5
7
9
12
15
18
TrafficType


Frequency of VisitorType
Count
0
6000
New_Visitor
Other
VisitorType


Frequency of Weekend
Count
0
6000
FALSE
TRUE
Weekend

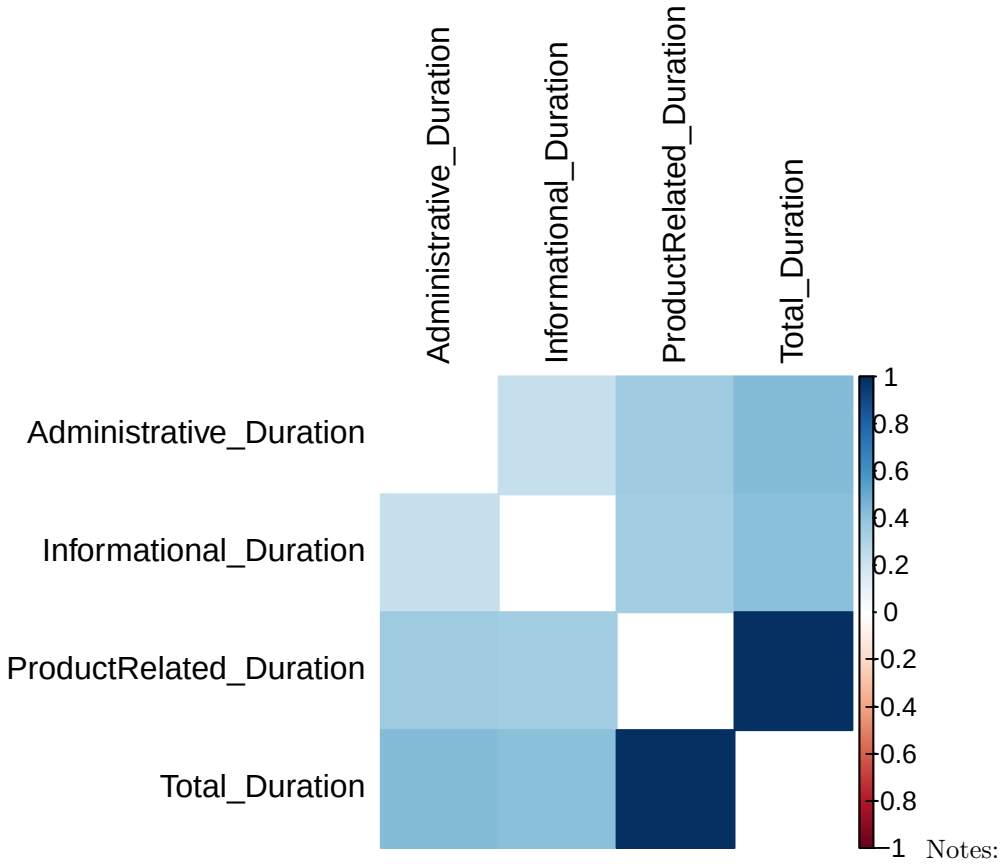

Frequency of Revenue
Count
0
6000
FALSE
TRUE
Revenue

Seeing if combining durations is useful

```
# Creating new dataset with total durations
shopping_uci_total_durations <- shopping_uci %>%
  mutate(Total_Duration = Administrative_Duration + Informational_Duration + ProductRelated_Duration)

shopping_uci_total_durations
```

```
## # A tibble: 12,330 x 19
##    Administrative Administrative_Duration Informational Informational_Duration
##             <dbl>                   <dbl>         <dbl>                   <dbl>
##  1              0                       0             0                       0
##  2              0                       0             0                       0
##  3              0                       0             0                       0
##  4              0                       0             0                       0
##  5              0                       0             0                       0
##  6              0                       0             0                       0
##  7              0                       0             0                       0
##  8              1                       0             0                       0
##  9              0                       0             0                       0
## 10              0                       0             0                       0
## # i 12,320 more rows
## # i 15 more variables: ProductRelated <dbl>, ProductRelated_Duration <dbl>,
## #   BounceRates <dbl>, ExitRates <dbl>, PageValues <dbl>, SpecialDay <dbl>,
## #   Month <fct>, OperatingSystems <fct>, Browser <fct>, Region <fct>,
## #   TrafficType <fct>, VisitorType <fct>, Weekend <fct>, Revenue <fct>,
## #   Total_Duration <dbl>
```

```
# Checking to see if total duration correlates with other individual durations
corr_values <- shopping_uci_total_durations %>%
  select(c(Administrative_Duration, Informational_Duration, ProductRelated_Duration, Total_Duration)) %
  cor()

corrplot(
  corr_values,
  diag = FALSE,
  method = "color",
  tl.col = "black"
)
```

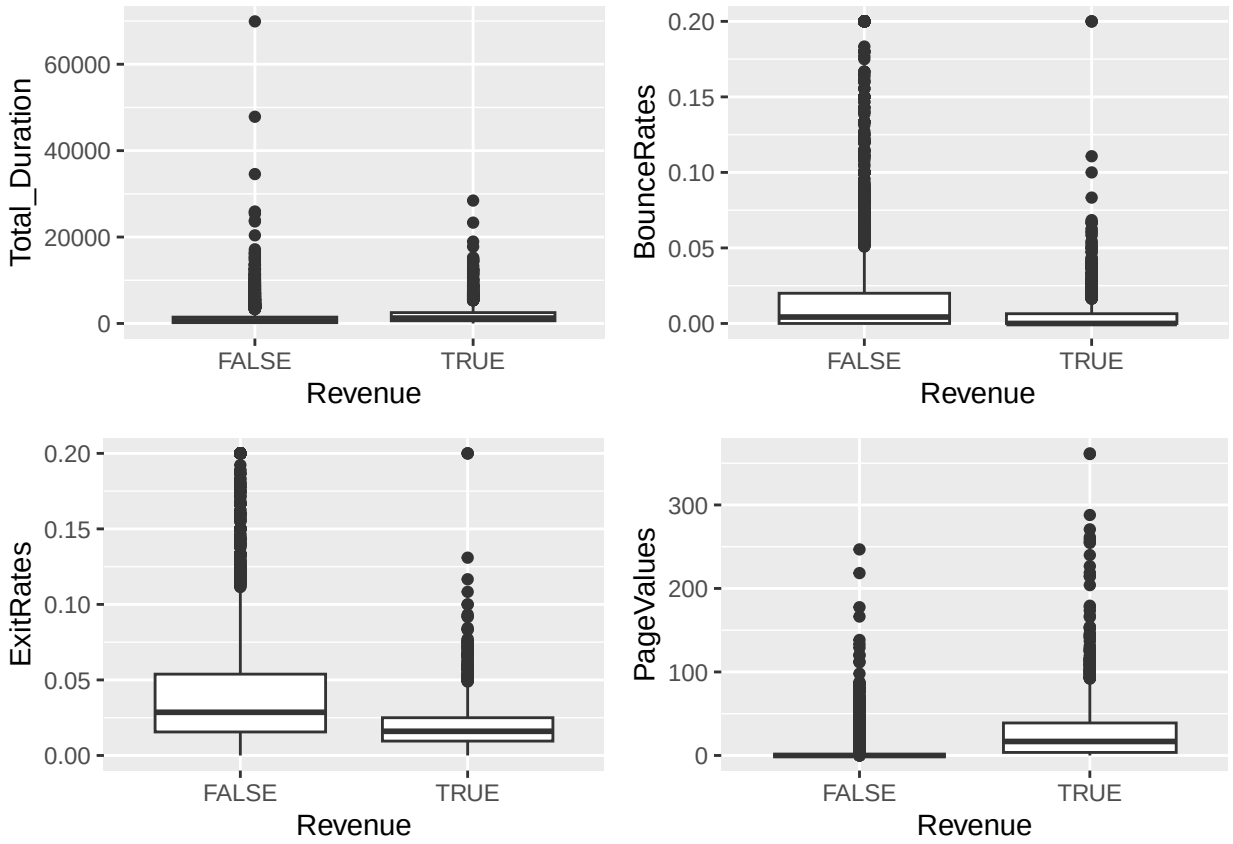


Notes:

- `Total_Duration` seems to correlate with the other individual durations, especially `ProductRelated_Duration`
- From this, it may be safe to sum up the individual durations into one variable, as it looks like we would not lose that much information if we perform this operation

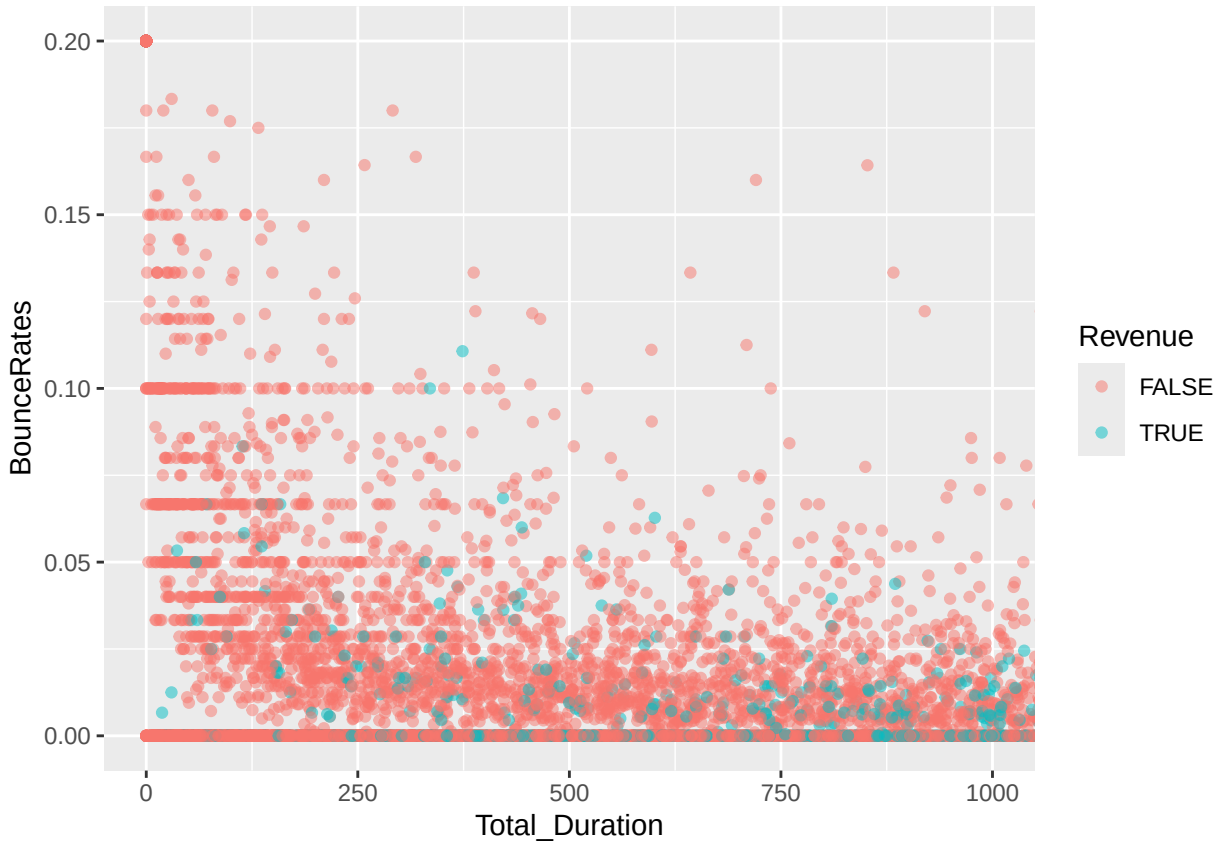
Multivariate analysis

```
# Box plots
total_duration <- ggplot(data = shopping_uci_total_durations, mapping = aes(x = Revenue, y = Total_Dura
  geom_boxplot()
bounce <- ggplot(data = shopping_uci_total_durations, mapping = aes(x = Revenue, y = BounceRates)) +
  geom_boxplot()
exit <- ggplot(data = shopping_uci_total_durations, mapping = aes(x = Revenue, y = ExitRates)) +
  geom_boxplot()
page <- ggplot(data = shopping_uci_total_durations, mapping = aes(x = Revenue, y = PageValues)) +
  geom_boxplot()

grid.arrange(total_duration, bounce, exit, page, ncol = 2)
```

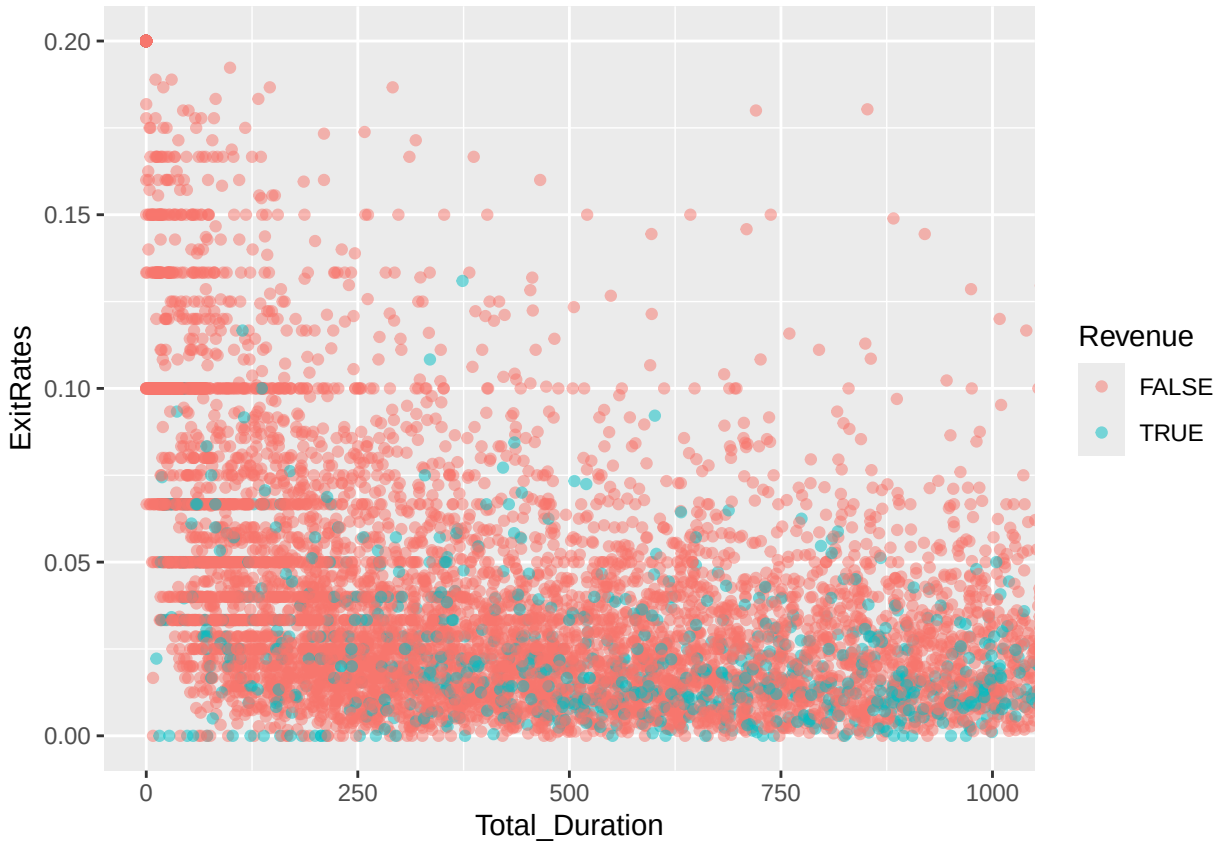


Nothing really useful here.

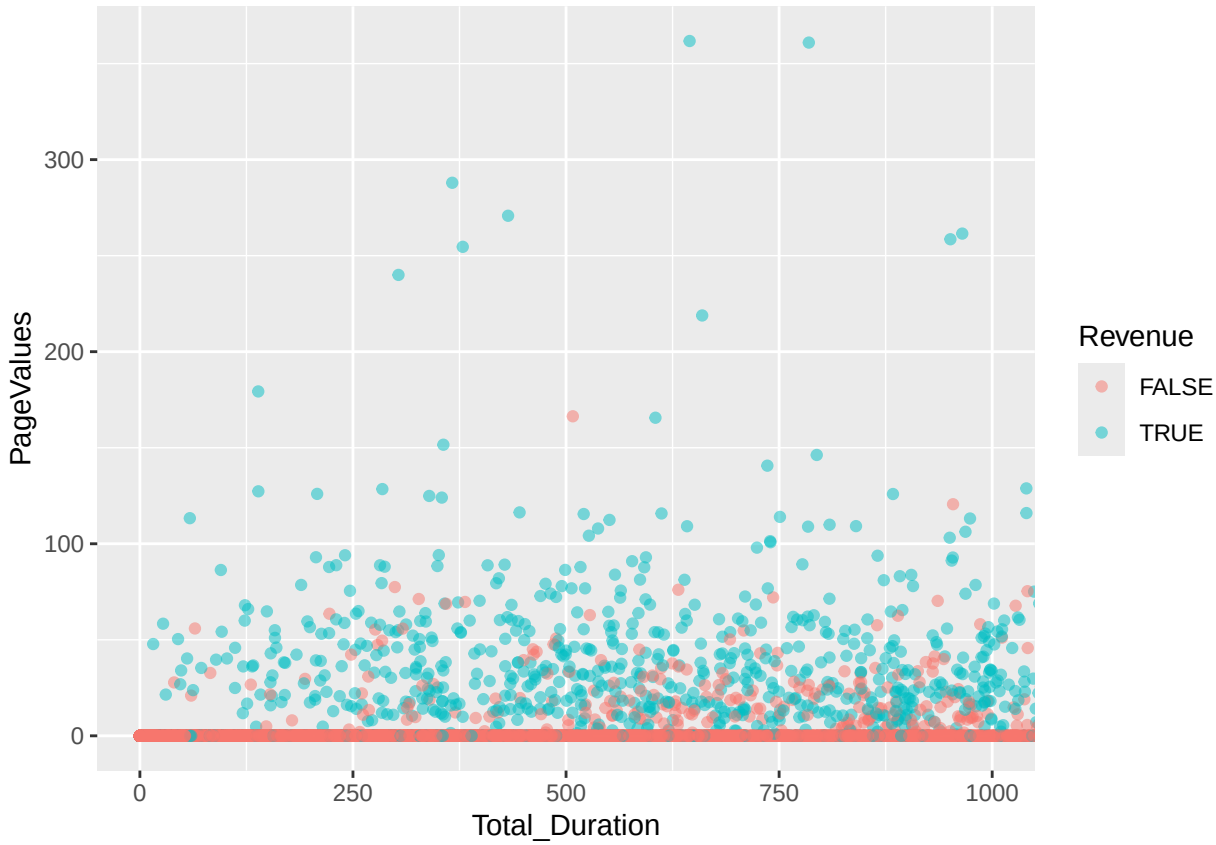
```
# Scatter plots
ggplot(data = shopping_uci_total_durations, mapping = aes(x = Total_Duration, y = BounceRates,
                                                          color = Revenue)) +
  geom_point(alpha = 0.5) +
  coord_cartesian(xlim = c(0, 1000))
```




```
ggplot(data = shopping_uci_total_durations, mapping = aes(x = Total_Duration, y = ExitRates,
                                                          color = Revenue)) +
  geom_point(alpha = 0.5) +
  coord_cartesian(xlim = c(0, 1000))
```

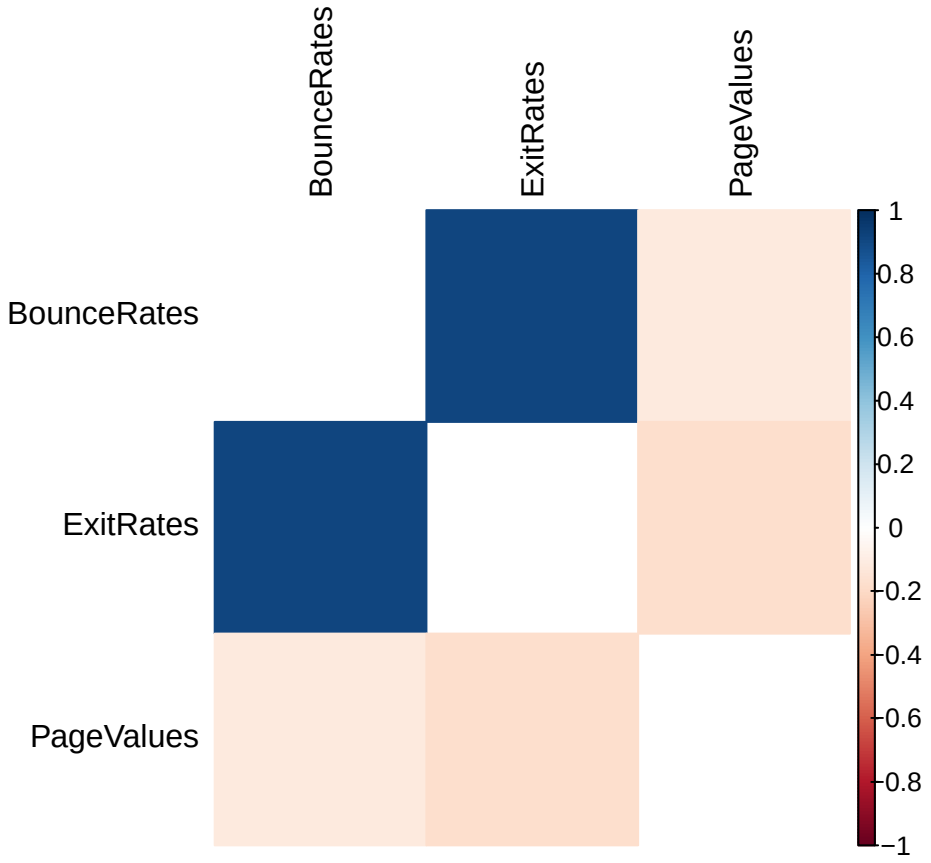



```
ggplot(data = shopping_uci_total_durations, mapping = aes(x = Total_Duration, y = PageValues,
                                                          color = Revenue)) +
  geom_point(alpha = 0.5) +
  coord_cartesian(xlim = c(0, 1000))
```



Nothing really useful can extracted from this as well

```
# Correlation plot between rates
corr_values <- shopping_uci_total_durations %>%
  select(c(BounceRates, ExitRates, PageValues)) %>%
  cor()

corrplot(
  corr_values,
  diag = FALSE,
  method = "color",
  tl.col = "black"
)
```


BounceRates
ExitRates
PageValues
BounceRates
ExitRates
PageValues
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1