Intro Data Science long project

Pedro Miguel de Sousa Magalhães

2022-01-14

Contents

Chapter 1

Introduction

1.1 Students commitment

> Declaro que o presente relatório é de minha autoria e não foi utilizado previamente noutro curso ou unidade curricular, desta ou de outra instituição. As referências a outros autores (afirmações,ideias, pensamentos) respeitam escrupulosamente as regras da atribuição, e encontramse devidamente indicadas no texto e nas referências bibliográficas, de acordo com as normas dereferenciação. Tenho consciência de que a prática de plágio e auto-plágio constitui um ilícito académico.

1.2 Software information

The R session information when building this project is has shown below:

```
sessionInfo()
```

```
## R version 3.5.3 (2019-03-11)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19042)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=Portuguese_Portugal.1252  LC_CTYPE=Portuguese_Portugal.1252
## [3] LC_MONETARY=Portuguese_Portugal.1252 LC_NUMERIC=C
## [5] LC_TIME=Portuguese_Portugal.1252
```

```
##
## attached base packages:
## [1] stats     graphics  grDevices datasets  utils     methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_3.5.3  magrittr_2.0.1  bookdown_0.24   fastmap_1.1.0
##  [5] htmltools_0.5.2 tools_3.5.3     rstudioapi_0.13 yaml_2.2.1
##  [9] stringi_1.7.6   rmarkdown_2.11  knitr_1.37      stringr_1.4.0
## [13] xfun_0.29       digest_0.6.29   rlang_0.4.12    renv_0.15.0
## [17] evaluate_0.14
```

Chapter 2

Business Understanding

2.1 Business objectives

The client is a e-commerce company operating on the high end market. Its only known sales channel is online. They wish to improve their customers experience, and their conversion, by using the information they actively collect from each touchpoint.

Main business goal: improve conversion How to achieve goal: understand which factors influence conversion so marketing strategy can be improved

2.2 Assess Situation

A Sample of customer sessions for given period of time is provided. There is no information regarding how session was defined. No information regarding user identification was provided and there for is not possible to use user as a perspective on the analysis or any information regarding acquisition journey. Therefore the project will focus solely on sessions, nonetheless it is important to point out the fact that a user can have several sessions which lead into a conversion and that can impact conversion strategies.

No special hardware and environment needs was identified

2.3 Data Mining Goals

Based on the Business Goals and the nature of the data available we can conclude this is a **binary classification problem with a focus on inference.** Therefore the following assumptions can be made about the expected output:

- The actual model contains valuable information to be used by the client. Therefore, black box models are less in line with the needs,
- The probability of conversion is not relevant,
- Each variable attribution is relevant

Throut this project we will address the following questions:

- Is there a relationship between a conversion and information available related to that session ?
- Which is the contribution of each of the variables to conversion?
- How accurately can we estimate the effect on conversion?
- Is there synergy among each session elements?
- Does a model surpass a naive baseline approach of assuming the most shown class?
- Does Data imbalance impact output?

2.4 Project Plan

The current project was executed with the following stages:

1. Explore data Analysis
2. Data transformations
3. Data preparation for modeling
4. Modeling
5. Conclusions

The following terms will be used during this project with the following meaning:

User: any unique IP which has reached the store. One individual can have more than one ip,

Client: a user that converted, this means, it bought from the shop,

Touchpoints: represents any interaction between the user and the online store of any sort,

Session: a period of time (normally of 30 min max) during which the user interacted with the shop. Every session starts with a touchpoint. Under some conditions depending on the website metrics collection a session can have more than one touchpoint.

Chapter 3

Data aquisition and understanding

Features by: - behavior - journey - segment - device and geo

```
library(tidyverse)
library(corrplot)
library(car)
library(tidymodels)
library(mice)
library(forecast)

source("scripts/eda_functions.r")
```

```
# import data
unbalanced_data <- read_csv("./data/train_full.csv") %>% select(-...1)
```

```
## New names:
## * `` -> ...1
```

```
## Rows: 95000 Columns: 21
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (9): session_id, plaform, segment, customer_type, device_group, visitor...
## dbl (12): ...1, has_listing, has_used_search, has_recommendation, has_add_to...
```

```
##
## iÃ Use `spec()` to retrieve the full column specification for this data.
## iÃ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
balanced_data <- read_csv("./data/train_balanced.csv") %>% select(-...1)
```

```
## New names:
## * `` -> ...1
```

```
## Rows: 16000 Columns: 21
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (9): session_id, plaform, segment, customer_type, device_group, visitor...
## dbl (12): ...1, has_listing, has_used_search, has_recommendation, has_add_to...
```

```
##
## iÂ Use `spec()` to retrieve the full column specification for this data.
## iÂ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# convert categorical and dummy to factor variable
unbalanced_data <- unbalanced_data %>%
  mutate_at(vars(
    !contains(c("duration", "view_qty", "unique_product_qty","unique_browse_designer_q
    ), ~ as.factor(.))

balanced_data <- balanced_data %>%
  mutate_at(vars(
    !contains(c("duration", "view_qty", "unique_product_qty","unique_browse_designer_q
    ), ~ as.factor(.))
```

3.1 Data description

Each row represents a unique session. The data available has the following nature and description given the role they play:

Dependent variable:

bought: categorical variable flagging if a order was made during that session. If true the value is 1 and zero undewise. Category available already has dummy. **This is the target or dependent variable of this project**.

Independent variable Features to use with the models

customer type: category with 2 levels, "prospect" if it hasn't purchased before and "customer" if it's a repeated buyer.

```
unique(unbalanced_data$customer_type)
```

```
## [1] prospect customer
## Levels: customer prospect
```

device group: category variable with 3 levels each representing the device source for each session.

```
unique(unbalanced_data$device_group)
```

```
## [1] Mobile Web App        Desktop
## Levels: App Desktop Mobile Web
```

visitor type: categorical value with 2 levels representing if a given user is a new or recurring user. It differs from customer type because it focus on visits and not actual conversion, therefore a returning user can be a prospect.

```
unique(unbalanced_data$visitor_type)
```

```
## [1] new       returning
## Levels: new returning
```

has_listing, has_used_search, has_recommendation, has_add_to_wishlist, has_add_to_bag: are all boolean variables representing key milestones on business customer journey. They are TRUE/ 1 if a given session includes that step.

duration: continuous variable representing the session duration in seconds

view_qty: discrete variable measuring the number of views during a session. Based on the information given we assume it measures the number of page views during a session

unique_product_qty: given the definition on view_qty refered before, it measures the number of page views on product pages

unique_browse_designer_qty: given the definition on view_qty refered before, it measures the number of page views on unique designer pages. The more the number represents a certain user researched a lot of desiners during the browsing session. Discrete variable

unique_browse_category_qty: given the definition on view_qty refered before, it measures the number of page views on unique category pages. The more the number represents a certain user researched a lot of products categories during the browsing session. Discrete variable

browser_name: categorical variable representing the browser used as source for each session. From the available information from the unbalanced data we have 52 different browsers present on this dataset.

```
unique(unbalanced_data$browser_name)
```

```
##  [1] Safari                  Chrome                  <NA>
##  [4] Instagram App           Mobile Safari UIWebView Android WebView
##  [7] Facebook App            Firefox                 Yandex Browser
## [10] Miui Browser            Google App              Opera
## [13] Edge                    Samsung Browser         HuaweiBrowser
## [16] Vivo Browser            Line App                Naver
## [19] WeChat App              Opera Mobile            DuckDuckGo Browser
## [22] Sogou Explorer          AliApp                  Silk
## [25] Apple Mail              Firefox for iOS         WKBrowser
## [28] UC Browser              HeyTapBrowser           Maxthon
## [31] Whale Browser           Snapchat                CM Browser
## [34] Weibo                   Default Browser         Tungsten Browser
## [37] QQBrowser               Ecosia                  Sleipnir
## [40] Android                 RDDocuments App         Coc Coc Browser
## [43] Meizu Browser           DareBoost Bot           Playstation Browser
## [46] Puffin                  Netease Music           Waterfox
## [49] Elements Browser        Iron                    Mail Master
## [52] Edge Mobile
## 51 Levels: AliApp Android Android WebView Apple Mail Chrome ... Yandex Browser
```

country: categorical variable containing the country of origin for each session.

```
unique(unbalanced_data$country)
```

```
##   [1] US   MX   RU   AU   IN   PT   KR   CL   GB   AE   CN   HK   DE   GR   PL
##  [16] KW   LI   BR   AT   HR   CA   IT   SA   TW   ZA   VN   JP   FR   RO   QA
##  [31] ME   BH   ID   PE   ES   BE   PK   LB   IE   KZ   BG   PH   NL   AR   IL
##  [46] SE   NZ   GE   DO   DK   BY   AF   MD   TH   MY   UA   LT   CH   EG   NO
##  [61] CO   SG   BA   RS   KH   HU   MO   IS   OM   AM   UY   MA   NG   AL   BB
##  [76] EE   DZ   MK   SK   LU   AO   TR   PA   EC   JO   IQ   BD   GH   CY   SI
##  [91] AZ   CZ   MU   KG   MT   BN   FI   TN   LV   CR   BS   SV   UZ   GT   AD
## [106] NP   HN   CM   KE   ET   AI   MC   GP   JM   <NA> MW   SN   VE   VI   CI
## [121] MR   ZM   AW   CG   LK   PR   GU   JE   GG   BM   SR   GI   TT   GY   BW
## [136] AQ   MN   MQ   HT   LC   SM   GF   KY   BO   SL   LA   IC   KV   VC   TC
## [151] BJ   MZ   LS   RE   NC   DM   MV   UG   TG   GL   PY   NI   GM
## 162 Levels: AD AE AF AI AL AM AO AQ AR AT AU AW AZ BA BB BD BE BG BH BJ ... ZM
```

Additional information not used as features

session_id: Unique identifier for each session. Each row represents a unique session

```
# test if a single session can have more than one row
length( unique(unbalanced_data$session_id) ) == nrow(unbalanced_data)
```

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

3.2 Explore data

```
unbalanced_chi <- unbalanced_data %>%
  select_if(is.factor) %>%
  map(function(x) chisq.test(x, unbalanced_data$bought))
```

```
## Warning in chisq.test(x, unbalanced_data$bought): Chi-squared approximation may
## be incorrect

## Warning in chisq.test(x, unbalanced_data$bought): Chi-squared approximation may
## be incorrect

## Warning in chisq.test(x, unbalanced_data$bought): Chi-squared approximation may
## be incorrect

## Warning in chisq.test(x, unbalanced_data$bought): Chi-squared approximation may
## be incorrect
```

```
balanced_chi <- balanced_data %>%
  select_if(is.factor) %>%
  map(function(x) chisq.test(x, balanced_data$bought))
```

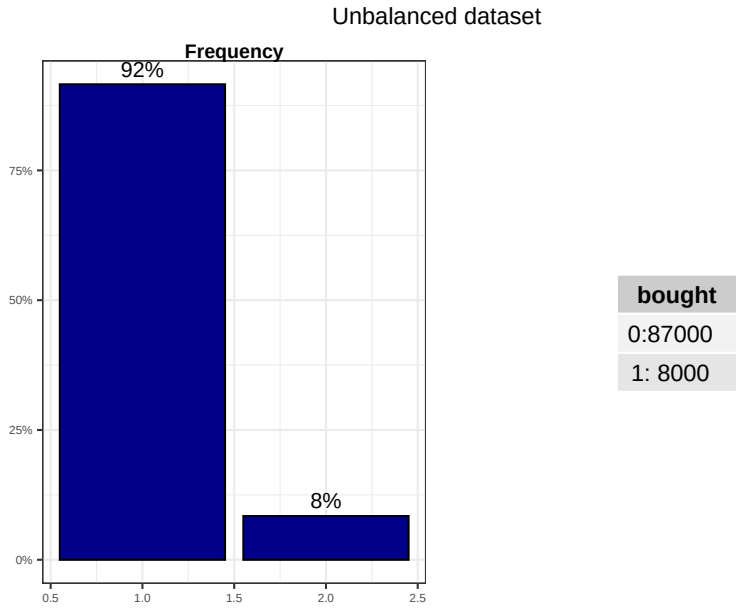
```
## Warning in chisq.test(x, balanced_data$bought): Chi-squared approximation may be
## incorrect

## Warning in chisq.test(x, balanced_data$bought): Chi-squared approximation may be
## incorrect

## Warning in chisq.test(x, balanced_data$bought): Chi-squared approximation may be
## incorrect

## Warning in chisq.test(x, balanced_data$bought): Chi-squared approximation may be
## incorrect
```

```
## Warning in chisq.test(x, balanced_data$bought): Chi-squared approximation may be
## incorrect
```

In this section we will explore the variables available using both unbalanced and balanced data. In special we will focus on the following questions:

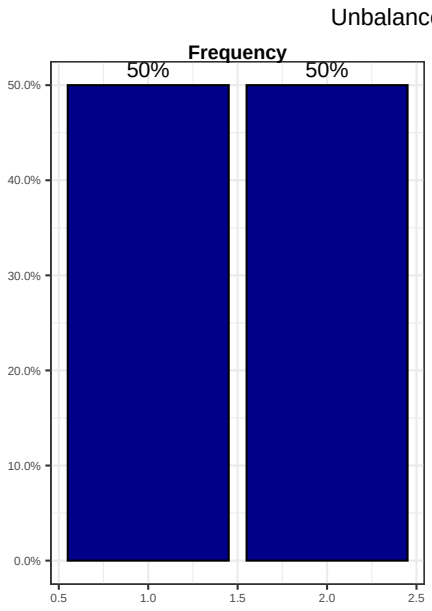
- What type of variation occurs within my variables?
- What type of covariation occurs between my variables?

3.2.1 Dependent variable: Bought



The "bought" variable is the target variable from this study and we can conclude the data is severely class inbalance towards no order which can affects modeling since it is biased towards the majority class. Has can be seen below this effect is corrected on the balanced dataset.

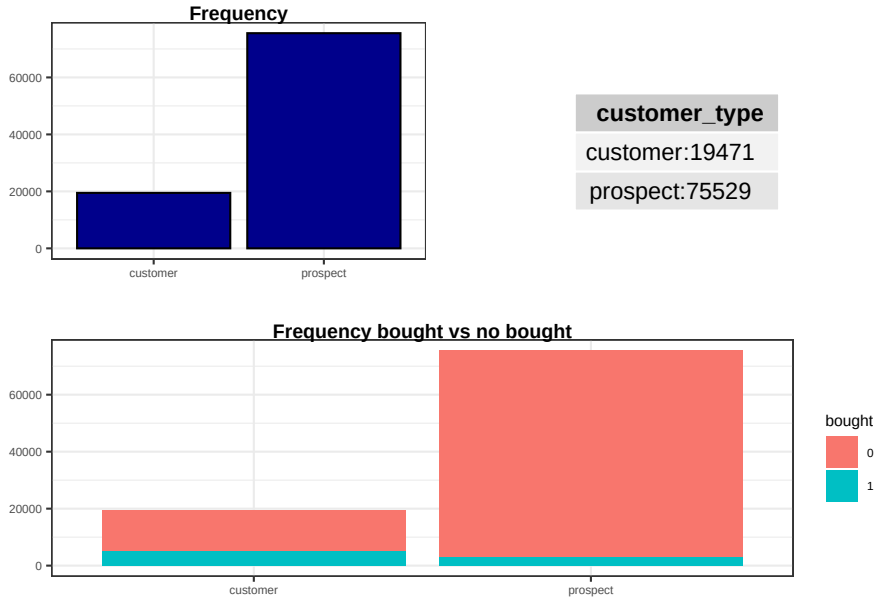
```
grid.draw(balanced_grid)
```



bought
0:8000
1:8000

3.2.2 Customer type

```
grid.draw(eda_unbalanced$customer_type)
```




```
grid.draw(eda_unbalanced$customer_type)
```

Customer type is a categorical variable with 2 levels. The current category in inbalanced with the majority of sessions being done by prospect clients. When

compared with target the inbalanced differs slightly being bought the majority class for customers implying a relationship between the 2.

The resulst from the χ^2 hipotesis test does not refuse the null hipothesys reinforcing the graphical analysis that a relationship might exist between this 2 variables that implies that recurrent customers buy more.

```
unbalanced_chi["customer_type"]
```

```
## $customer_type
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  x and unbalanced_data$bought
## X-squared = 9750.7, df = 1, p-value < 2.2e-16
```

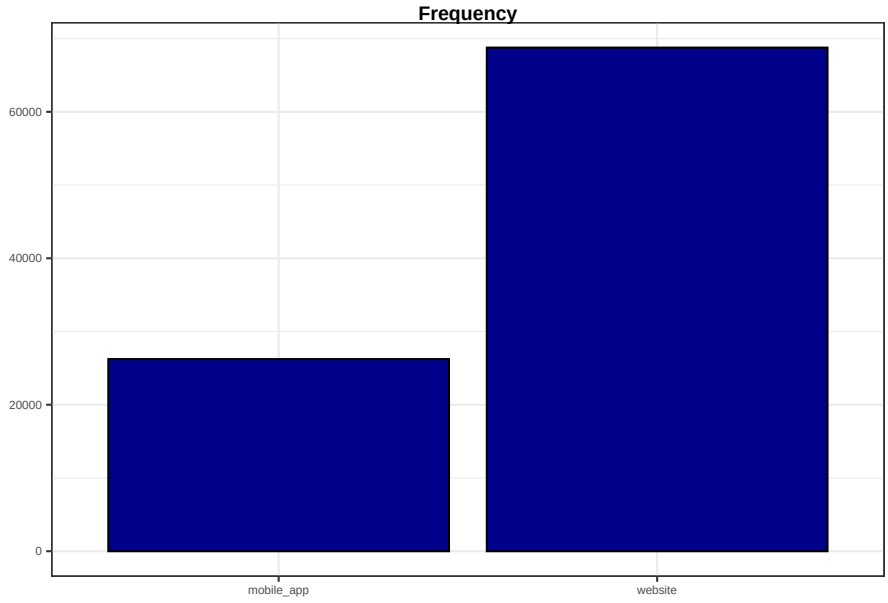
```
balanced_chi["customer_type"]
```

```
## $customer_type
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  x and balanced_data$bought
## X-squared = 3590, df = 1, p-value < 2.2e-16
```

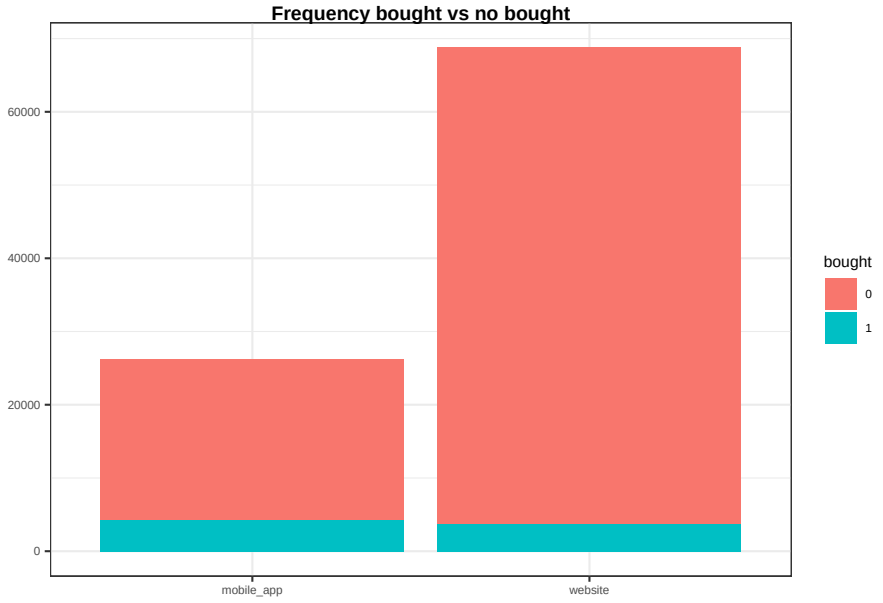
3.2.3 Platform

```
grid.draw(eda_unbalanced$platform)
```

```
ggplot(unbalanced_data, aes(x = plaform)) +
    geom_bar(fill="darkblue", color="black") +
    theme_masterDS() +
    labs(
      x = "",
      y = "",
      title = "Frequency"
    )
```




```
ggplot(unbalanced_data, aes( x = plaform, fill = bought )) +
    geom_bar() +
    theme_masterDS() +
    labs(
      x = "",
      y = "",
      title = "Frequency bought vs no bought"
    )
```




```
prop.table(table(unbalanced_data[["plaform"]]))
```

```
##
## mobile_app    website
##  0.2761263  0.7238737
```

```
unbalanced_chi["plaform"]
```

```
## $plaform
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  x and unbalanced_data$bought
## X-squared = 2896.4, df = 1, p-value < 2.2e-16
```

```
grid.draw(eda_balanced$platform)
```

```
prop.table(table(balanced_data[["plaform"]]))
```

```
##
## mobile_app    website
##    0.39125    0.60875
```

```
balanced_chi["plaform"]
```

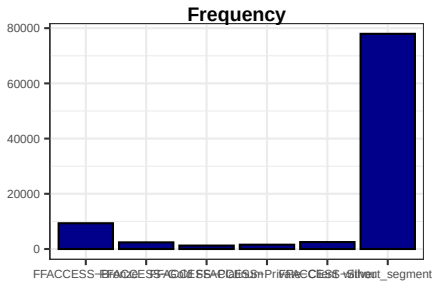
```
## $plaform
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  x and balanced_data$bought
## X-squared = 1360.5, df = 1, p-value < 2.2e-16
```

The graphical analysis show us that the majority of sessions (around 72% on the unbalanced dataset and 61% on the balanced) were accessed through the website. Despite the inbalance between platforms we notice that the amount of conversion is comparable suggesting implying a higher conversion rate on mobile app compared to website.

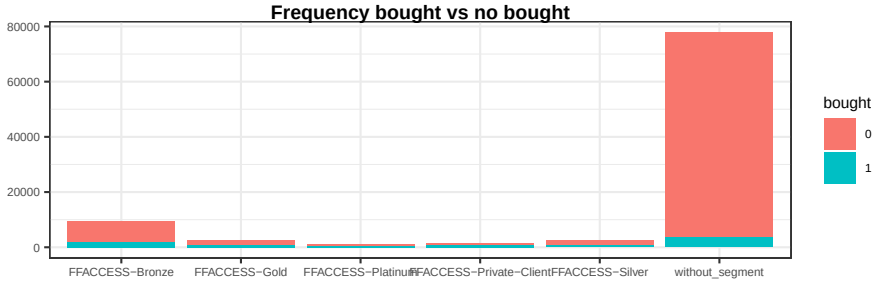
The χ^2 for both unbalanced and balanced datasets do not allow for the rejection of the null hypothesis implying the existence of a degree of linear regression between the 2 variables.

3.2.4 Segment

```
grid.draw(eda_unbalanced$segment)
```



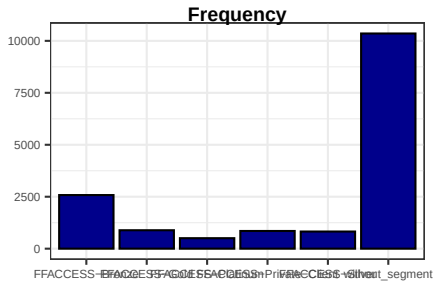
segment	
FFACCESS−Bronze	: 9331
FFACCESS−Gold	: 2405
FFACCESS−Platinum	: 1240
FFACCESS−Private−Client	: 1564
FFACCESS−Silver	: 2511
without_segment	:77949



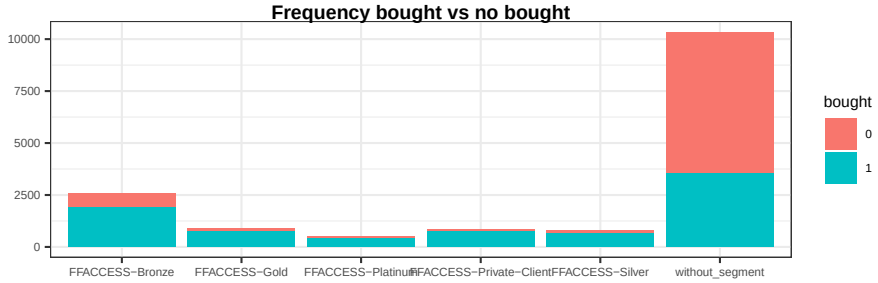
```
unbalanced_chi["segment"]
```

```
## $segment
##
##  Pearson's Chi-squared test
##
## data:  x and unbalanced_data$bought
## X-squared = 10203, df = 5, p-value < 2.2e-16
```

```
grid.draw(eda_balanced$segment)
```



segment	
FFACCESS−Bronze	: 2579
FFACCESS−Gold	: 887
FFACCESS−Platinum	: 504
FFACCESS−Private−Client	: 851
FFACCESS−Silver	: 825
without_segment	:10354




```
balanced_chi["segment"]
```

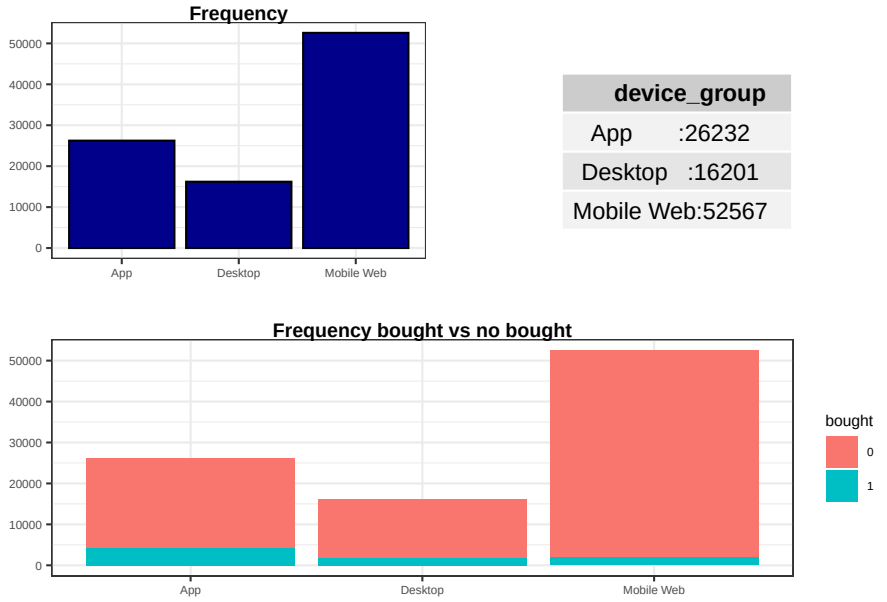
```
## $segment
##
##  Pearson's Chi-squared test
##
## data:  x and balanced_data$bought
## X-squared = 3071, df = 5, p-value < 2.2e-16
```

The graphical analysis shows that the great majority of sessions were executed by users not belonging to any segment. Does not point to any relationship.

The χ^2 test for both unbalanced and balanced datasets do not allow for the rejection of the null hypothesis implying the existence of a degree of linear regression between the 2 variables.

3.2.5 Device Group

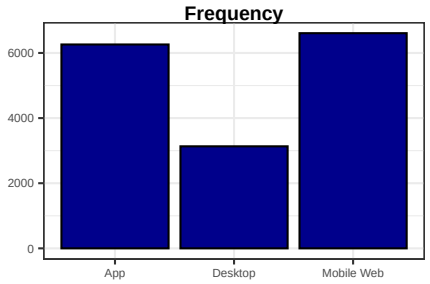
```
grid.draw(eda_unbalanced$device_group)
```



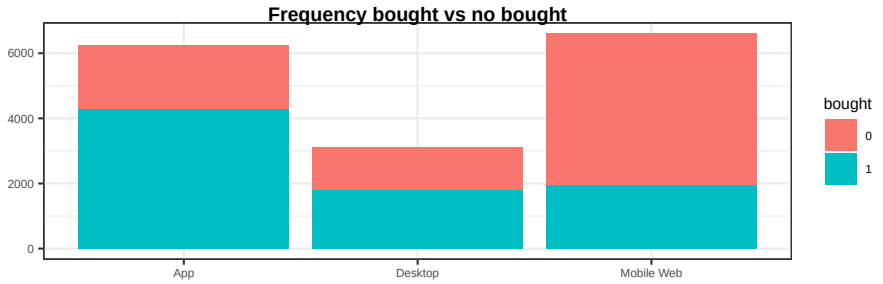

```
unbalanced_chi["device_group"]
```

```
## $device_group
##
##  Pearson's Chi-squared test
##
## data:  x and unbalanced_data$bought
## X-squared = 3755.4, df = 2, p-value < 2.2e-16
```

```
grid.draw(eda_balanced$device_group)
```



device_group	
App	:6260
Desktop	:3133
Mobile Web	:6607




```
balanced_chi["device_group"]
```

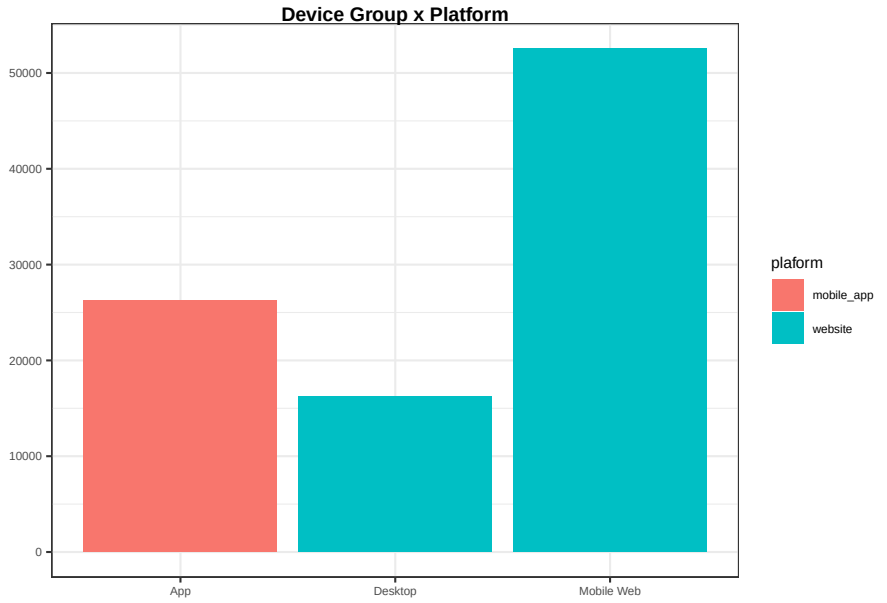
```
## $device_group
##
##  Pearson's Chi-squared test
##
## data:  x and balanced_data$bought
## X-squared = 2003.4, df = 2, p-value < 2.2e-16
```

The graphical analysis of the data tells us that the majority of sessions have been done using Mobile Web platform, although, when taking into account the actual conversion it suggests that a higher conversion rate exists on application or desktop than compared with Majority class.

This same conclusions can be extracted from the balanced dataset although the differences are not as evident. The χ^2 tests for both dataset leads into refusing the null hypothesis so it suggests that a degree of relation ship exists between both variables.

```
ggplot(unbalanced_data, aes(x = device_group, fill = plaform)) +
  geom_bar(stat = "count") +
  theme_masterDS() +
    labs(
      x = "",
```

```
    y = "",
    title = "Device Group x Platform"
  )
```



This dataset includes a variable named platform that suggest that most sessions were done using website. At first glance it seems counter intuitive that most access be done through website and Mobile web but the above plot show that in reality most users opt to access using the mobile version of the website instead of the app.

The χ^2 test between this 2 variables confirms what we could already suspect from visual inspection, both variable seem to have a degree of relation leading to not refuting the null hiopothesis of dependeny. This implies interaction (or sinergy on marketing terms) between features which can impact how modeling efforts specially for models dependent on linear transformations (least squares regression) as is the case of logit.

During the modeling step one might consider removing one of the variables or generate a new compound feature.

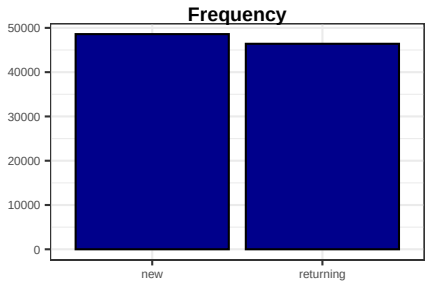
```
chisq.test(as.character(unbalanced_data$plaform), as.character(unbalanced_data$device_
```

```
##
##  Pearson's Chi-squared test
##
```

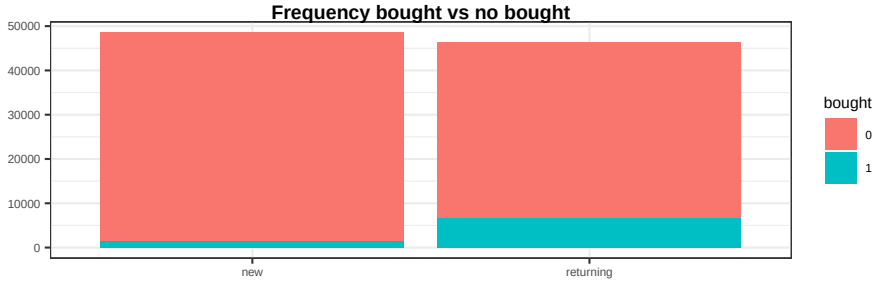
```
## data:  as.character(unbalanced_data$plaform) and as.character(unbalanced_data$device_group)
## X-squared = 95000, df = 2, p-value < 2.2e-16
```

3.2.6 Visitor type

```
grid.draw(eda_unbalanced$visitor_type)
```



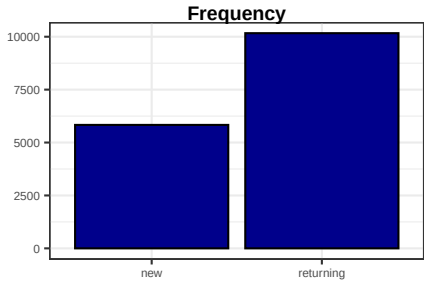
visitor_type	
new	:48592
returning	:46408



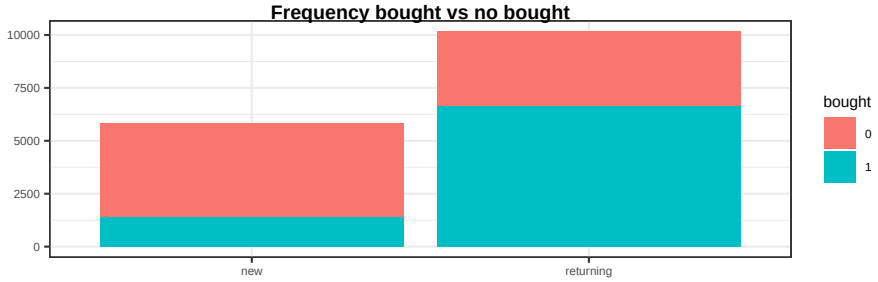

```
unbalanced_chi["visitor_type"]
```

```
## $visitor_type
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  x and unbalanced_data$bought
## X-squared = 4001.3, df = 1, p-value < 2.2e-16
```

```
grid.draw(eda_balanced$visitor_type)
```



visitor_type	
new	: 5833
returning	:10167




```
balanced_chi["visitor_type"]
```

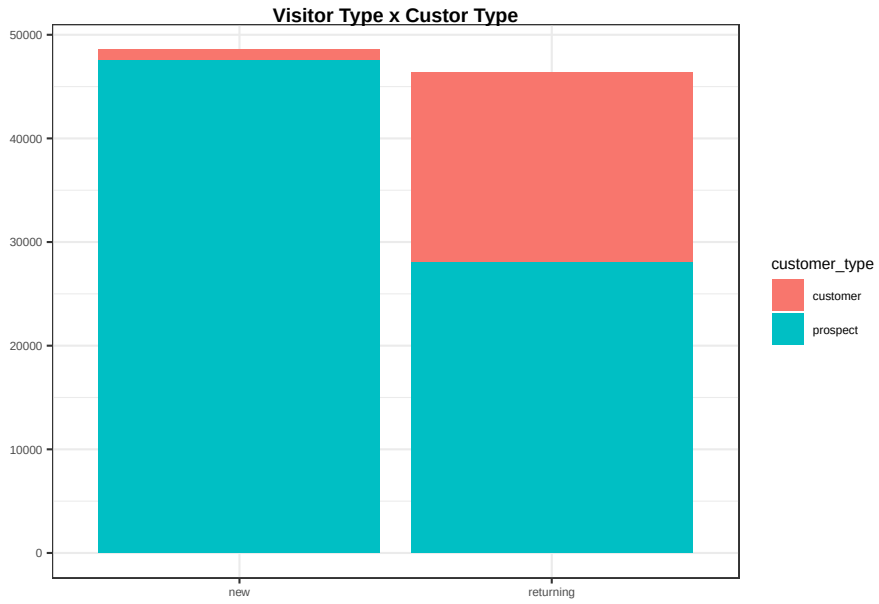
```
## $visitor_type
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  x and balanced_data$bought
## X-squared = 2529.6, df = 1, p-value < 2.2e-16
```

This feature focus on the Visitors. The unbalanced data available shows almost a 50% split, situations that changes on the balanced dataset which has a inbalance towards returning visitors. The graphical analysis suggests a higher conversion rate for returning visitor that for new one, suggesting that continuous visits (engagment) plays a role in conversion.

The current dataset has information regarding customer type crossed with visitor type can provide us with interesting information

```
ggplot(unbalanced_data, aes(x = visitor_type, fill = customer_type)) +
  geom_bar(stat = "count") +
  theme_masterDS() +
    labs(
      x = "",
      y = "",
```

```
    title = "Visitor Type x Custor Type"
  )
```

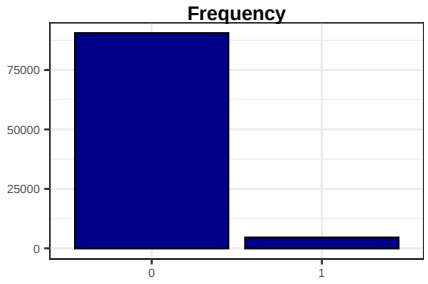


Has expected only a fraction of new visitors actually buy on the first session hinting that the conversion journey is longer than one session, meaning than several visits are needed before a first conversion. We don't have enough information to conclude about the number of sessions needed (journey lenght) and neither the session index given a time window (eg: the current session is the #3 in the last 28 days) which is know to have a impact on conversion.

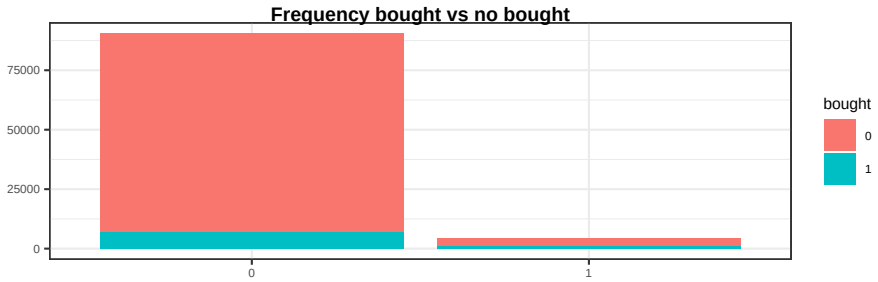
3.2.7 Journey milestones

(has_listings) > has_add_to_bag

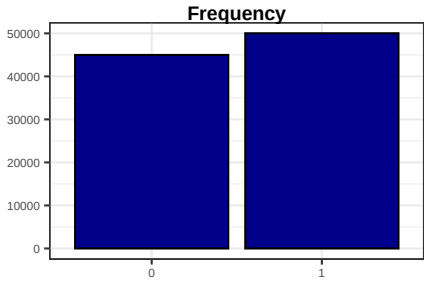
```
grid.draw(eda_unbalanced$has_add_to_wishlist)
```



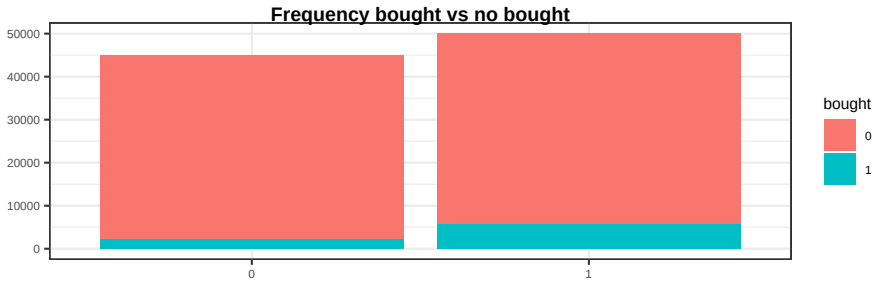
has_add_to_wishlist
0:90468
1: 4532



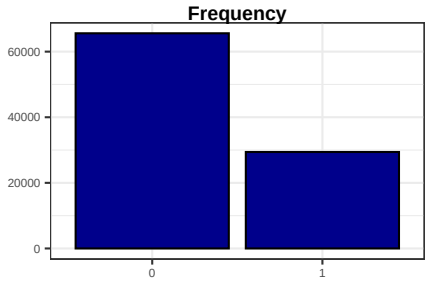

```
grid.draw(eda_unbalanced$has_listing)
```



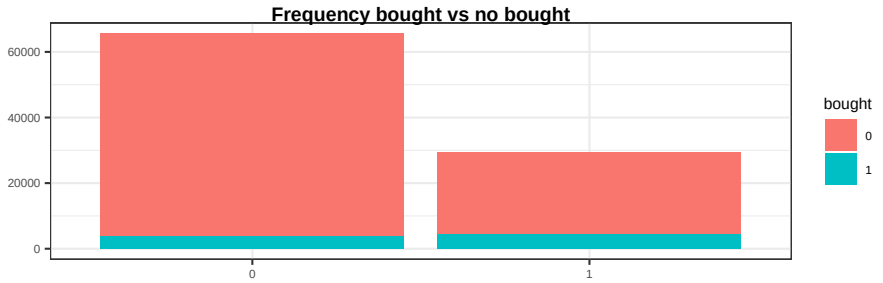
has_listing
0:44979
1:50021



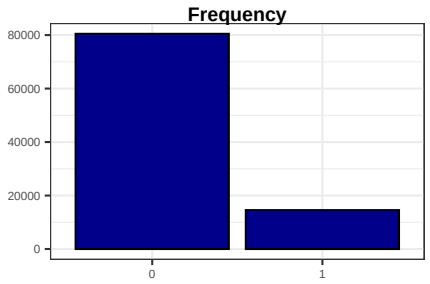
```
grid.draw(eda_unbalanced$has_recommendation)
```



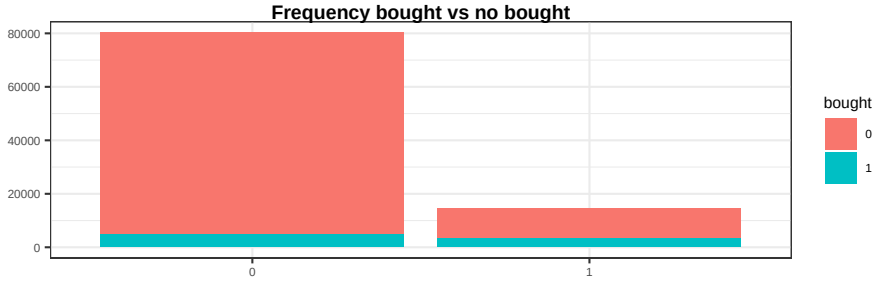
has_recommendation
0:65594
1:29406



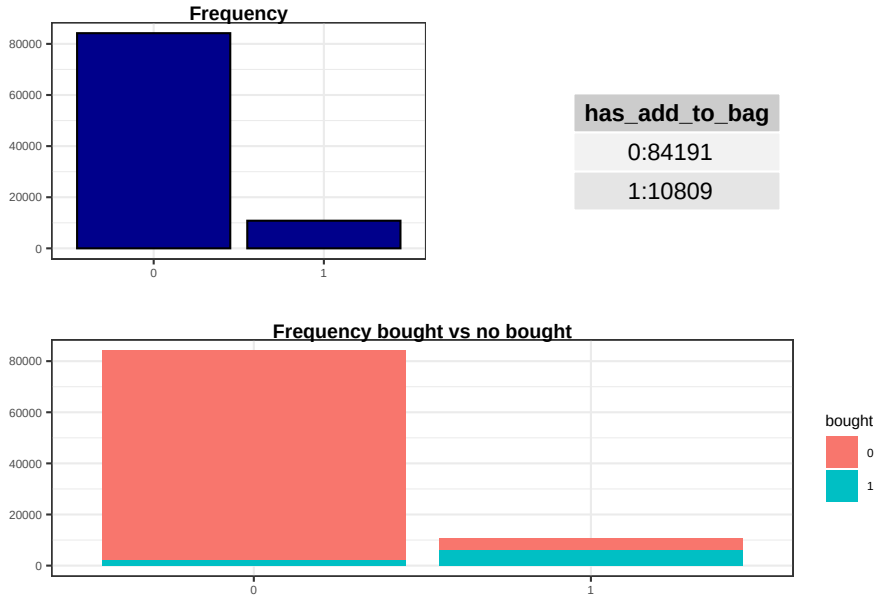

```
grid.draw(eda_unbalanced$has_used_search)
```



has_used_search
0:80466
1:14534



```
grid.draw(eda_unbalanced$has_add_to_bag)
```



This features are plotted together because they give insights into the journey a user made inside the website. We can conclude over a conversion funnel but it makes sense to explore the interactions between then while modeling because its known that normally strong effects exist between then (a user had a recommendation and added to the bag might be together a strong signal for conversion).

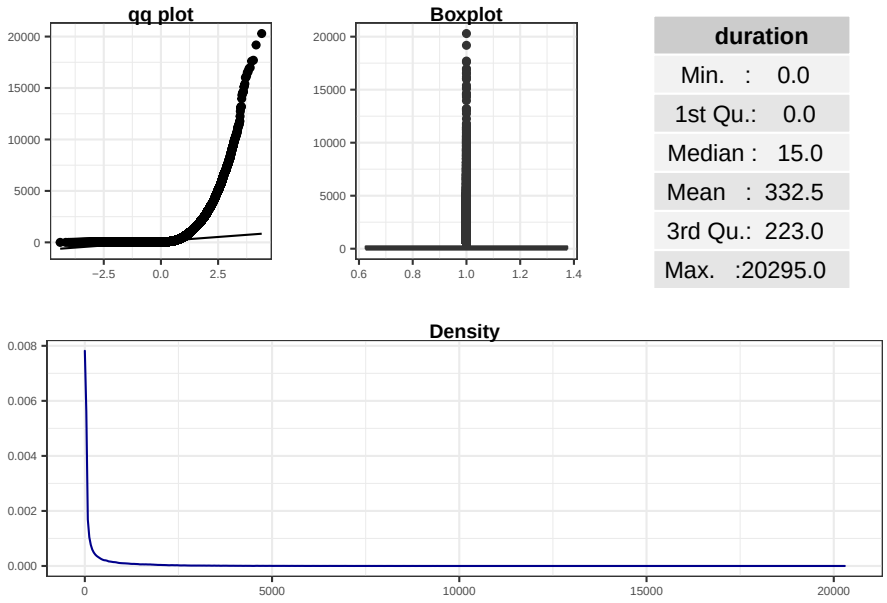
The graphical analysis already provides a important insight given the business objectives. From the unbalanced data available we can conclude that around of 46% shopping carts are lost on that session. That raises a question of how are this recovered (example on a next session) or if this means that all this sales are lost right at the end of the sales funnel.

```
prop.table(table(unbalanced_data[unbalanced_data$has_add_to_bag == 1,]$bought, unbalanc
```

```
##         0         1
## 0.4500879 0.5499121
```

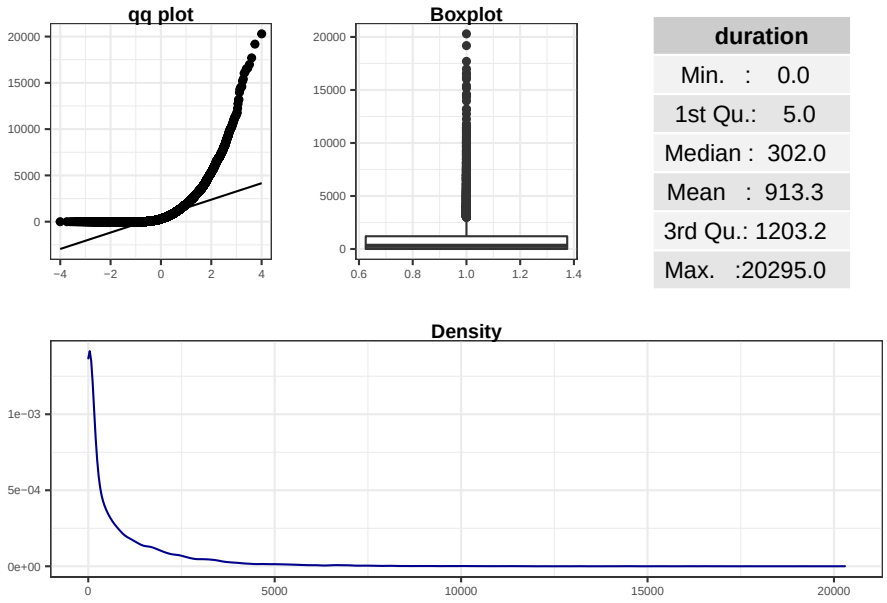
3.2.8 Duration

```
grid.draw(eda_unbalanced$duration)
```



duration
Min. : 0.0
1st Qu.: 0.0
Median : 15.0
Mean : 332.5
3rd Qu.: 223.0
Max. :20295.0

```
grid.draw(eda_balanced$duration)
```



duration
Min. : 0.0
1st Qu.: 5.0
Median : 302.0
Mean : 913.3
3rd Qu.: 1203.2
Max. :20295.0

Both unbalanced and balanced datasets show a left skewed distribution for duration. The Distance between the Mean and Median provides a good view of data. This extrem values have a big influence impacting the gaussianity of the distribution.

Normally web analytics software limit session duration between 25 to 30 min. It is assumed that beyond that point either the user is idle or there is a error on the collection. For periods of engagement superior to that cap normally a new session is started. So a 1hour of engagement would signify 2 simultaneous sessions.

In the case of our project we have no context regarding session duration. Therefore, we have no referencial from business which would lead us to remove.

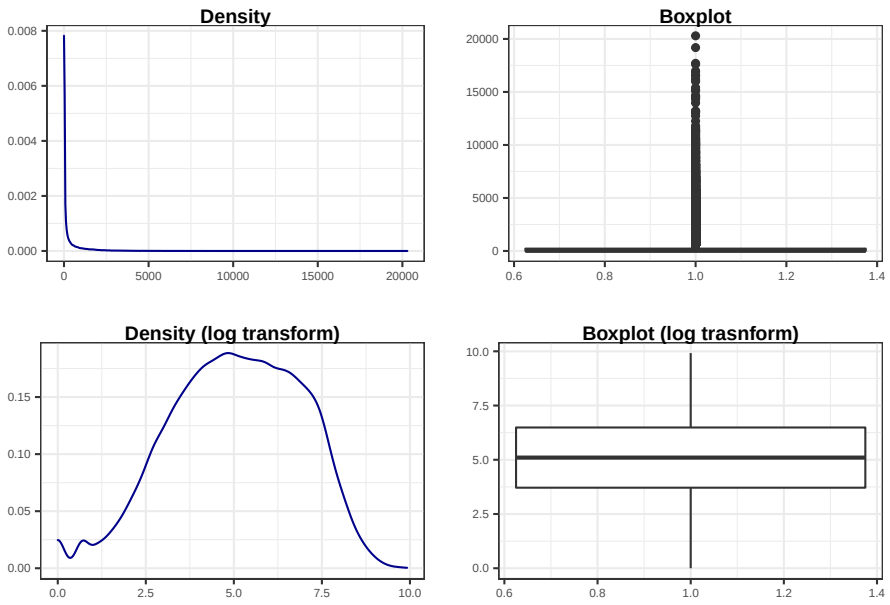
From the graphical view we can see that the variable distribution is left skewed whcih can cause us problems during modeling and impacts our outlier analysis. We will log transform this variable.

```
up <- quantile(unbalanced_data$duration, 0.75) + 3 * IQR(unbalanced_data$duration)

unbalanced_outlier_clean <- unbalanced_data %>%
  filter(duration < up)

  density <- ggplot(unbalanced_data, aes(x = duration)) +
    geom_density(color="darkblue") +
    theme_masterDS() +
    labs(
      x = "",
      y = "",
      title = "Density"
    )

  boxplot <- ggplot(unbalanced_data, aes(x = 1, y = duration)) +
    geom_boxplot() +
    theme_masterDS()+
    labs(
      x = "",
      y="",
      title = "Boxplot"
    )

  density_log <- ggplot(unbalanced_data, aes(x = log(duration))) +
    geom_density(color="darkblue") +
    theme_masterDS() +
    labs(
      x = "",
      y = "",
```

```
      title = "Density (log transform)"
    )

  boxplot_log <- ggplot(unbalanced_data, aes(x = 1, y = log(duration))) +
    geom_boxplot() +
    theme_masterDS()+
    labs(
      x = "",
      y="",
      title = "Boxplot (log trasnform)"
    )

grid.arrange(density, boxplot, density_log, boxplot_log,  layout_matrix = rbind(c(1,2),c(3,4)))
```

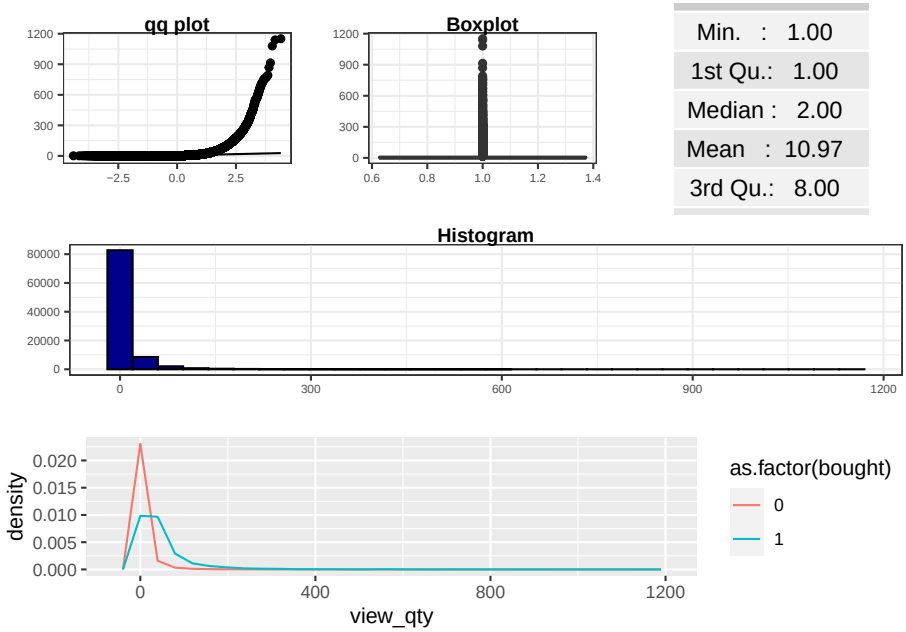
```
## Warning: Removed 41417 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 41417 rows containing non-finite values (stat_boxplot).
```

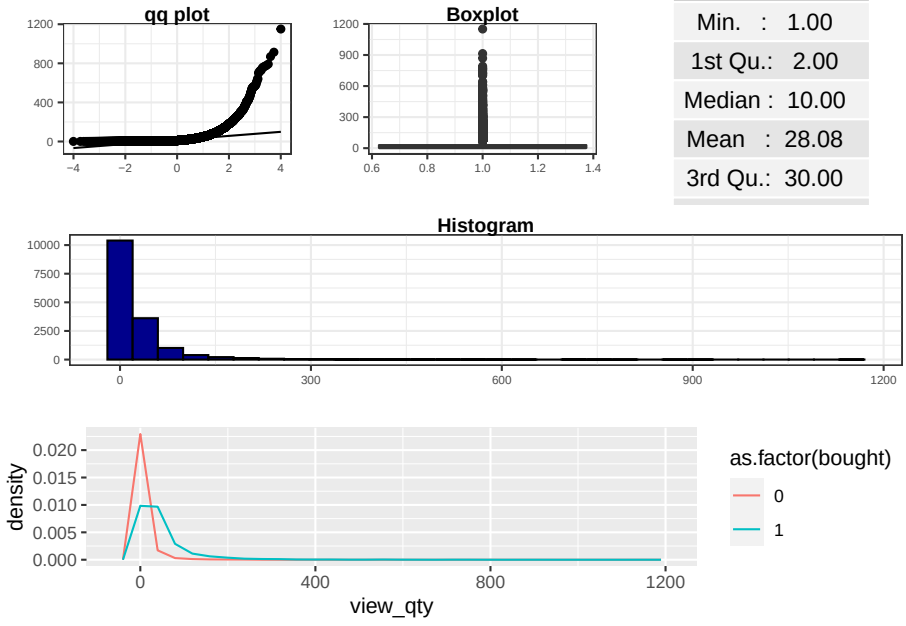


3.2.9 View quantity

```
grid.draw(eda_unbalanced$view_qty)
```



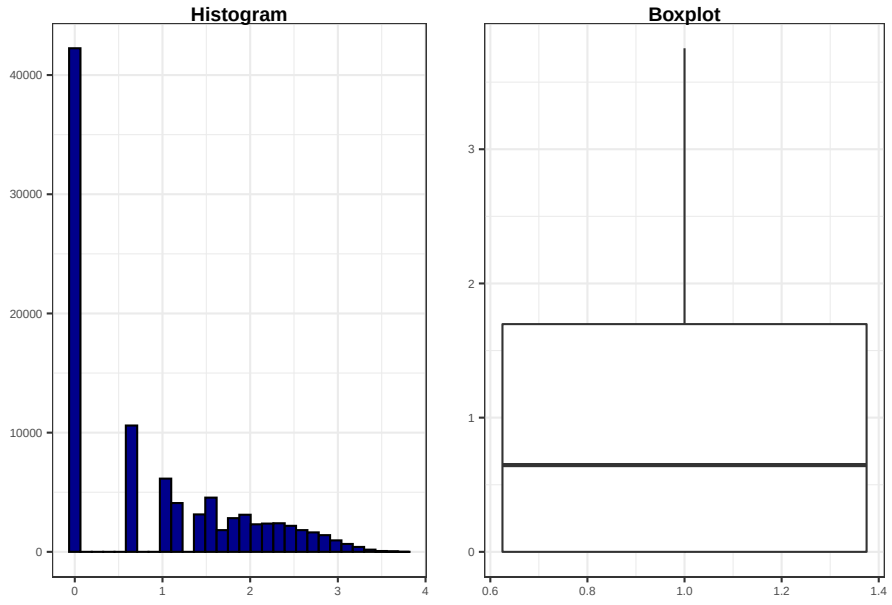

```
grid.draw(eda_balanced$view_qty)
```



left skwed apply log transform

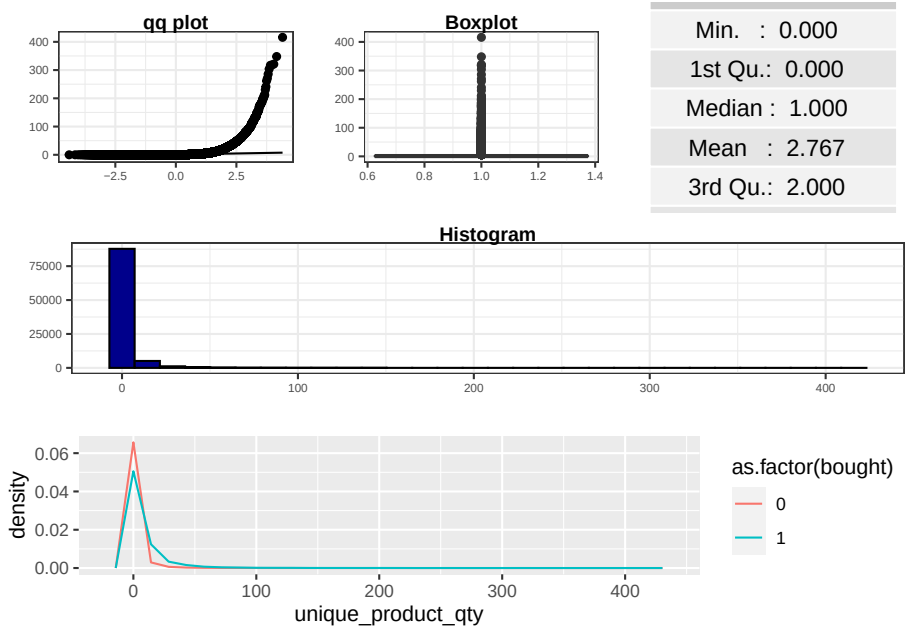
```
 lambda <- BoxCox.lambda(unbalanced_data$view_qty, method = "guerrero")

histogram_log <- ggplot(unbalanced_data, aes(x = bcPower(view_qty, lambda = lambda))) +
    geom_histogram(fill="darkblue", color="black") +
    theme_masterDS() +
    labs(
      x = "",
      y = "",
      title = "Histogram"
    )

  boxplot_log <- ggplot(unbalanced_data, aes(x = 1, y = bcPower(view_qty, lambda = lambda))) +
    geom_boxplot() +
    theme_masterDS()+
    labs(
      x = "",
      y="",
      title = "Boxplot"
    )

  grid.arrange(histogram_log, boxplot_log, nrow = 1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

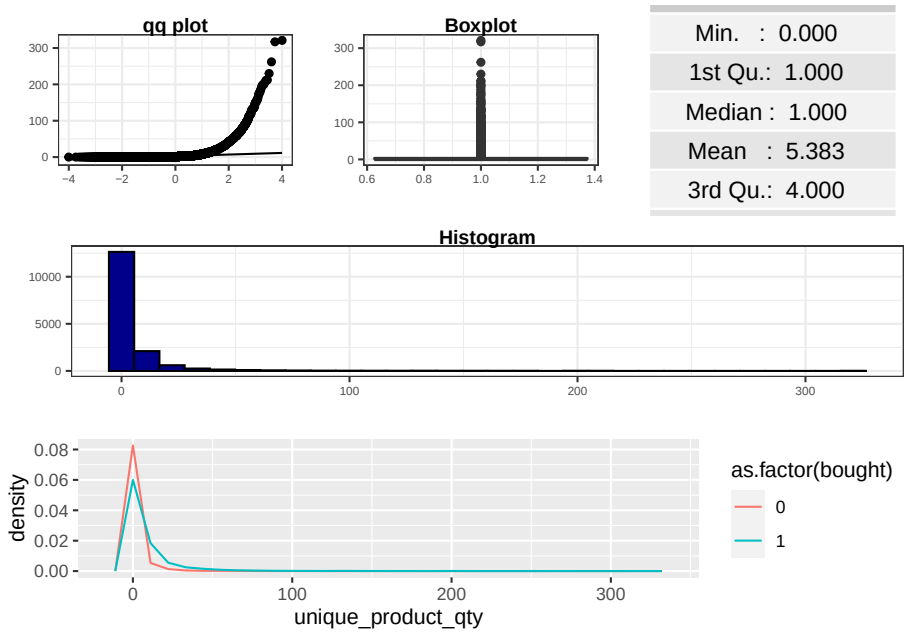


3.2.10 Unique product qty

```
grid.draw(eda_unbalanced$unique_product_qty)
```




```
grid.draw(eda_balanced$unique_product_qty)
```



left skwed apply log transform

```
lambda <- BoxCox.lambda(unbalanced_data$unique_product_qty, method = "guerrero")
```

```
## Warning in guerrero(x, lower, upper): Guerrero's method for selecting a Box-Cox
## parameter (lambda) is given for strictly positive data.
```

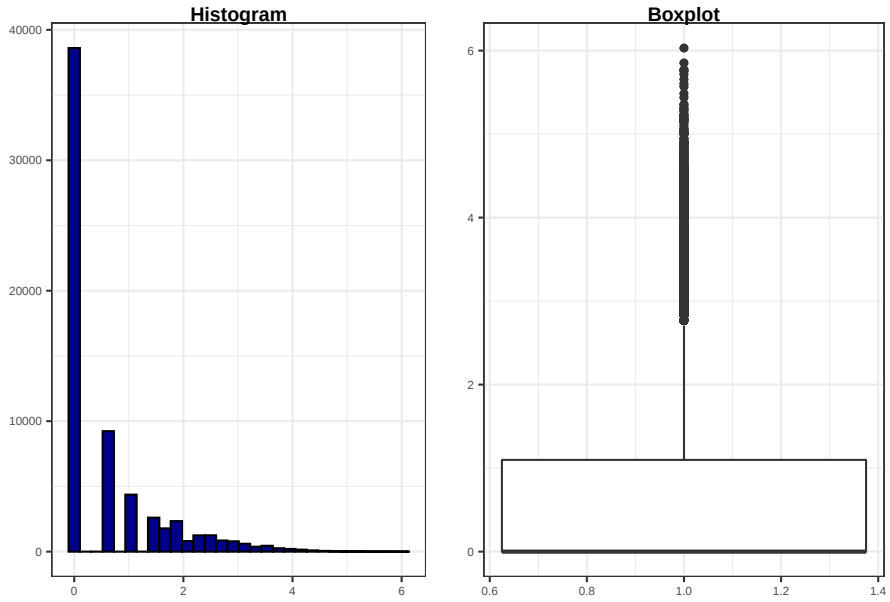
```
histogram_log <- ggplot(unbalanced_data, aes(x = log(unique_product_qty))) +
    geom_histogram(fill="darkblue", color="black") +
    theme_masterDS() +
    labs(
      x = "",
      y = "",
      title = "Histogram"
    )

  boxplot_log <- ggplot(unbalanced_data, aes(x = 1, y = log(unique_product_qty))) +
    geom_boxplot() +
    theme_masterDS()+
    labs(
      x = "",
      y="",
      title = "Boxplot"
    )
```

```
  grid.arrange(histogram_log, boxplot_log, nrow = 1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

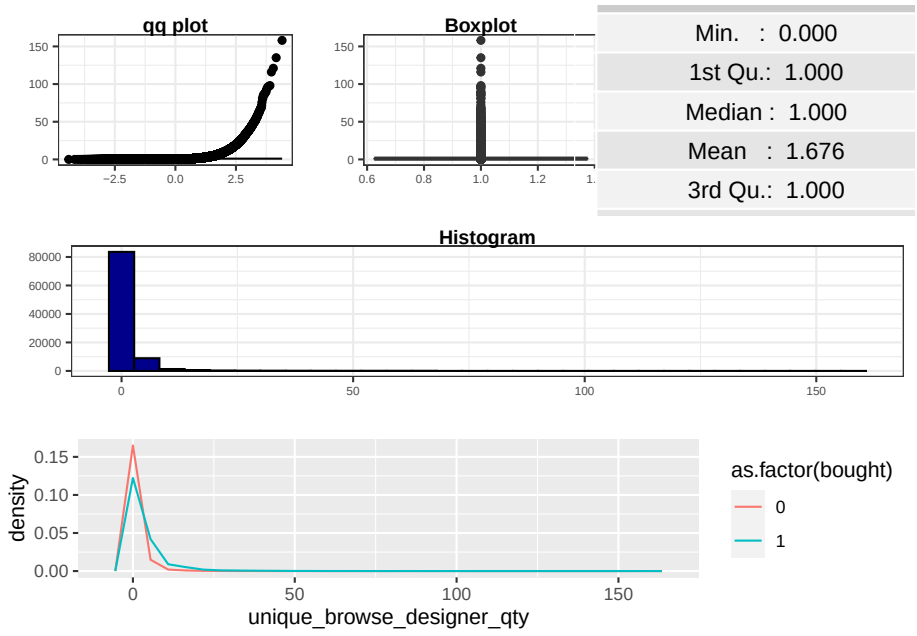
```
## Warning: Removed 28829 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 28829 rows containing non-finite values (stat_boxplot).
```

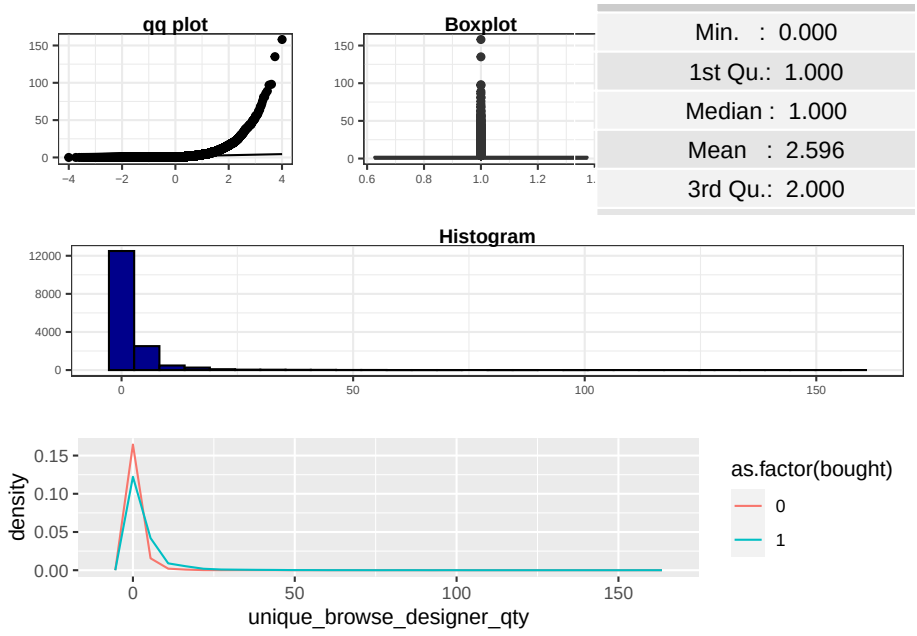


3.2.11 Unique browse designer quantity

```
grid.draw(eda_unbalanced$unique_browse_designer_qty)
```




```
grid.draw(eda_balanced$unique_browse_designer_qty)
```



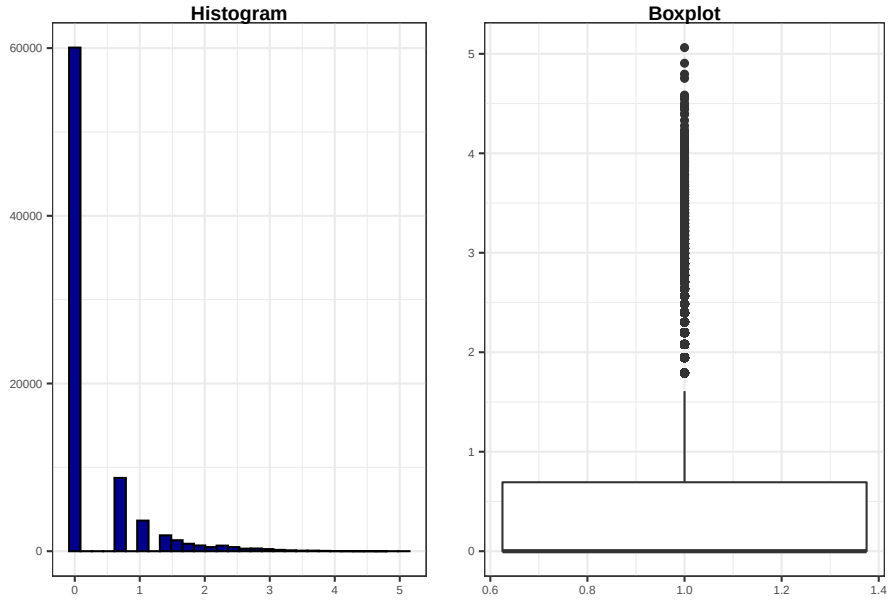
left skwed apply log transform

```
histogram_log <- ggplot(unbalanced_data, aes(x = log(unique_browse_designer_qty))) +
   geom_histogram(fill="darkblue", color="black") +
   theme_masterDS() +
   labs(
     x = "",
     y = "",
     title = "Histogram"
   )

 boxplot_log <- ggplot(unbalanced_data, aes(x = 1, y = log(unique_browse_designer_qty)
   geom_boxplot() +
   theme_masterDS()+
   labs(
     x = "",
     y="",
     title = "Boxplot"
   )

 grid.arrange(histogram_log, boxplot_log, nrow = 1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

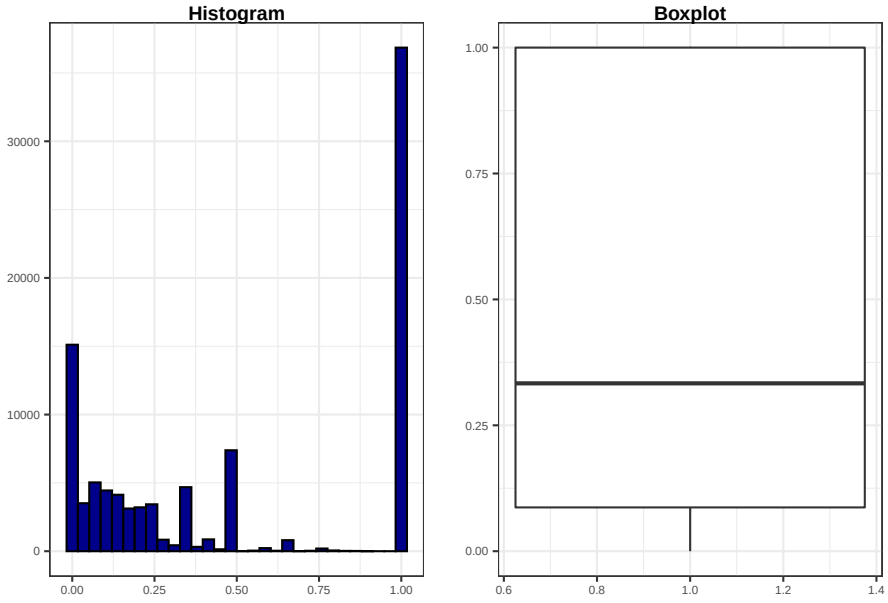
```
## Warning: Removed 14778 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 14778 rows containing non-finite values (stat_boxplot).
```



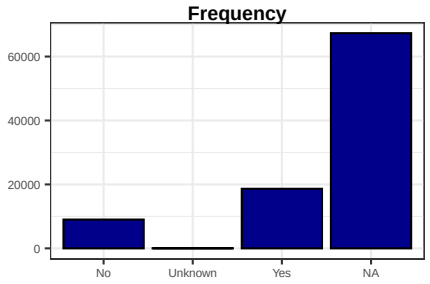

```
histogram_log <- ggplot(unbalanced_data, aes(x = unique_browse_designer_qty/view_qty)) +
   geom_histogram(fill="darkblue", color="black") +
   theme_masterDS() +
   labs(
     x = "",
     y = "",
     title = "Histogram"
   )

 boxplot_log <- ggplot(unbalanced_data, aes(x = 1, y = unique_browse_designer_qty/view_qty)) +
   geom_boxplot() +
   theme_masterDS()+
   labs(
     x = "",
     y="",
     title = "Boxplot"
   )

 grid.arrange(histogram_log, boxplot_log, nrow = 1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

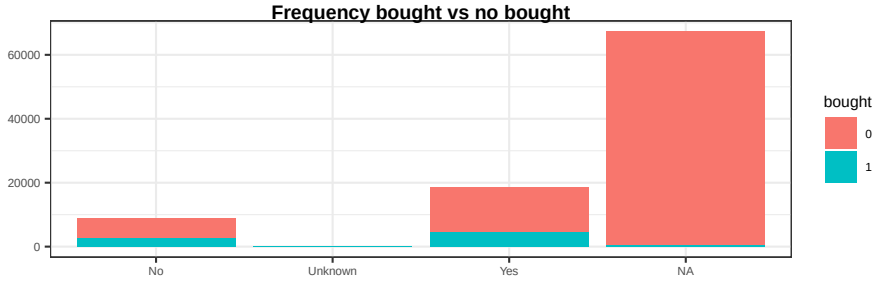


3.2.12 Is subscribed

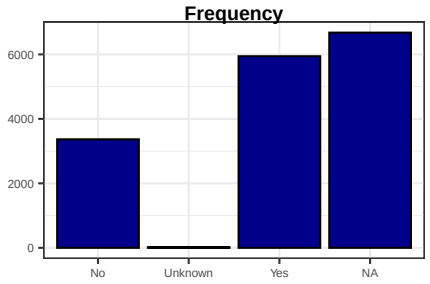
```
grid.draw(eda_unbalanced$is_subscribed)
```



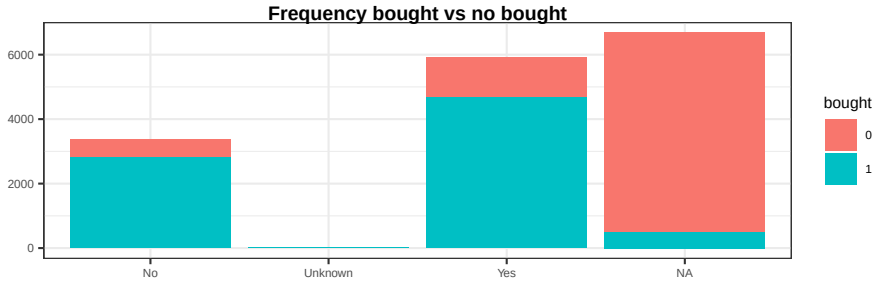
is_subscribed	
No	: 9004
Unknown:	43
Yes	:18640
NA's	:67313




```
grid.draw(eda_balanced$is_subscribed)
```



is_subscribed	
No	:3365
Unknown:	18
Yes	:5942
NA's	:6675



For both datasets the class majority is NA. From the information given we cannot conclude that we can remove this feature from the model, and given the

number of observations affected we will look for imputation alternatives.

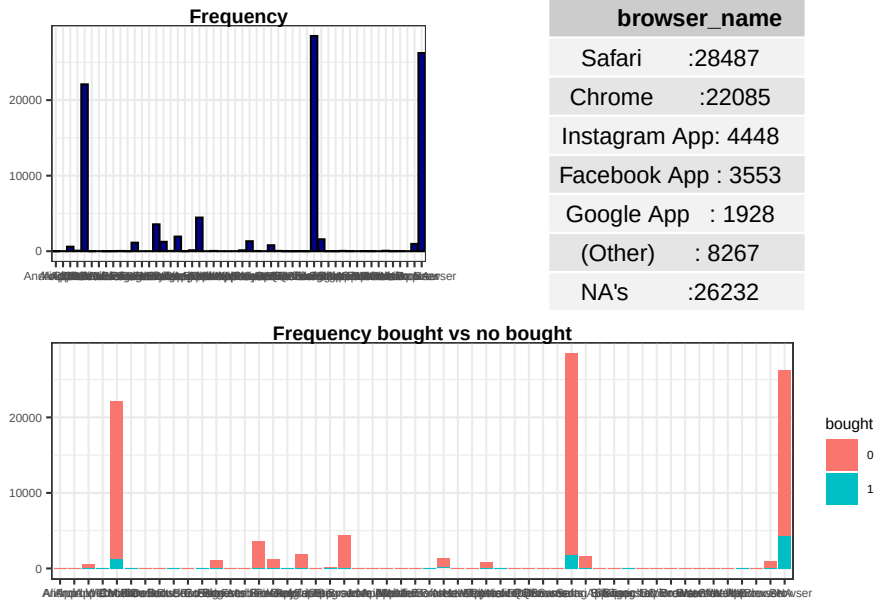
If a particular variable is having more missing values that rest of the variables in the dataset, and, if by removing that one variable you can save many observations. I would, then, suggest to remove that particular variable, unless it is a really important predictor that makes a lot of business sense. It is a matter of deciding between the importance of the variable and losing out on a number of observation.

Given that its a categorical variable we could replace with the mode, and that would mean that all NA would become Yes, but given the size of missing values this can return a huge impact. We will explore statitical imputation using knn.

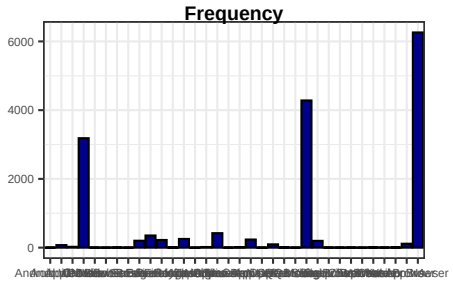
DMwR::knnImputation uses k-Nearest Neighbours approach to impute missing values. What kNN imputation does in simpler terms is as follows: For every observation to be imputed, it identifies 'k' closest observations based on the euclidean distance and computes the weighted average (weighted based on distance) of these 'k' obs.

3.2.13 Browser name

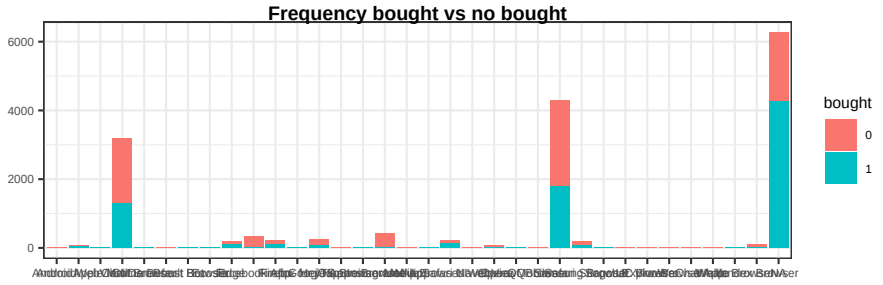
```
grid.draw(eda_unbalanced$browser_name)
```



```
grid.draw(eda_balanced$browser_name)
```

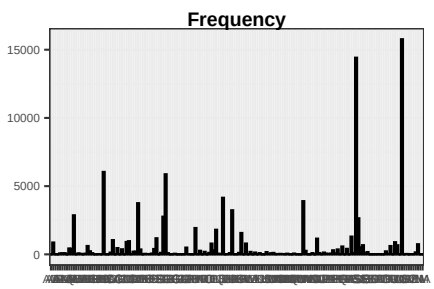


browser_name	
Safari	:4282
Chrome	:3184
Instagram App	: 421
Facebook App	: 353
Google App	: 253
(Other)	:1247
NA's	:6260

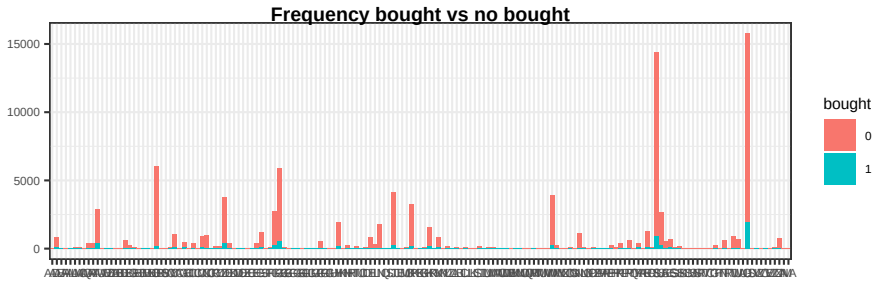


3.2.14 Country

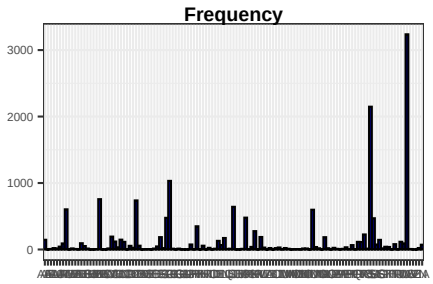
```
grid.draw(eda_unbalanced$country)
```



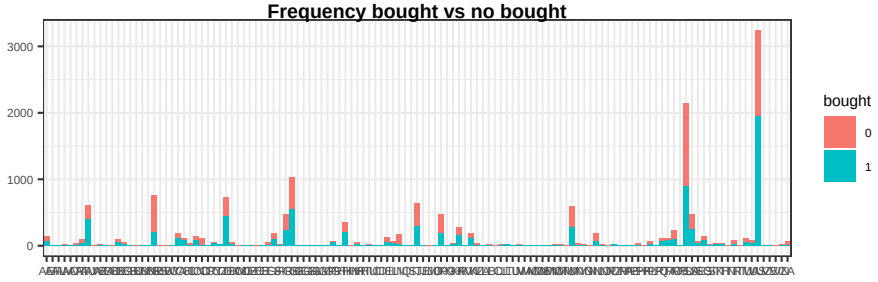
country	
US	:15772
RU	:14418
BR	: 6046
GB	: 5874
IT	: 4152
(Other)	:48734
NA's	: 4




```
grid.draw(eda_balanced$country)
```



country	
US	:3237
RU	:2148
GB	:1032
BR	: 757
DE	: 738
IT	: 642
(Other)	:7446



remove rows withna

```
correct_data <- select(unbalanced_data, -session_id)
summary(correct_data)
```

```
##        plaform                          segment       customer_type
##  mobile_app:26232   FFACCESS-Bronze        : 9331   customer:19471
##  website   :68768   FFACCESS-Gold          : 2405   prospect:75529
##                     FFACCESS-Platinum      : 1240
##                     FFACCESS-Private-Client: 1564
##                     FFACCESS-Silver        : 2511
##                     without_segment        :77949
##
##      device_group      visitor_type   has_listing has_used_search
##  App       :26232   new      :48592   0:44979     0:80466
##  Desktop   :16201   returning:46408   1:50021     1:14534
##  Mobile Web:52567
##
##
##
##
##  has_recommendation has_add_to_wishlist has_add_to_bag    duration
##  0:65594            0:90468             0:84191        Min.   :    0.0
##  1:29406            1: 4532             1:10809        1st Qu.:    0.0
##                                                        Median :   15.0
##                                                        Mean   :  332.5
##                                                        3rd Qu.:  223.0
##                                                        Max.   :20295.0
##
##     view_qty       unique_product_qty unique_browse_designer_qty
##  Min.   :   1.00   Min.   :  0.000    Min.   :  0.000
##  1st Qu.:   1.00   1st Qu.:  0.000    1st Qu.:  1.000
##  Median :   2.00   Median :  1.000    Median :  1.000
##  Mean   :  10.97   Mean   :  2.767    Mean   :  1.676
##  3rd Qu.:   8.00   3rd Qu.:  2.000    3rd Qu.:  1.000
##  Max.   :1151.00   Max.   :416.000    Max.   :158.000
##
##  unique_browse_category_qty is_subscribed          browser_name
##  1      :53094              No     : 9004   Safari       :28487
##  0      :22498              Unknown:   43   Chrome       :22085
##  2      :11476              Yes    :18640   Instagram App: 4448
##  3      : 3834              NA's   :67313   Facebook App : 3553
##  4      : 1795                              Google App   : 1928
##  5      :  929                              (Other)      : 8267
##  (Other): 1374                              NA's         :26232
##     country      bought
##  US     :15772   0:87000
```

```
##  RU     :14418   1: 8000
##  BR     : 6046
##  GB     : 5874
##  IT     : 4152
##  (Other):48734
##  NA's   :    4
```

```
rec <- recipe(bought ~ ., data = correct_data)

ratio_recipe <- rec %>%
  step_impute_knn(all_predictors(), neighbors = 3)
ratio_recipe2 <- prep(ratio_recipe, training = unbalanced_data, verbose = TRUE, retain
```

```
## oper 1 step impute knn [training]
## The retained training set is ~ 8.36 Mb  in memory.
```

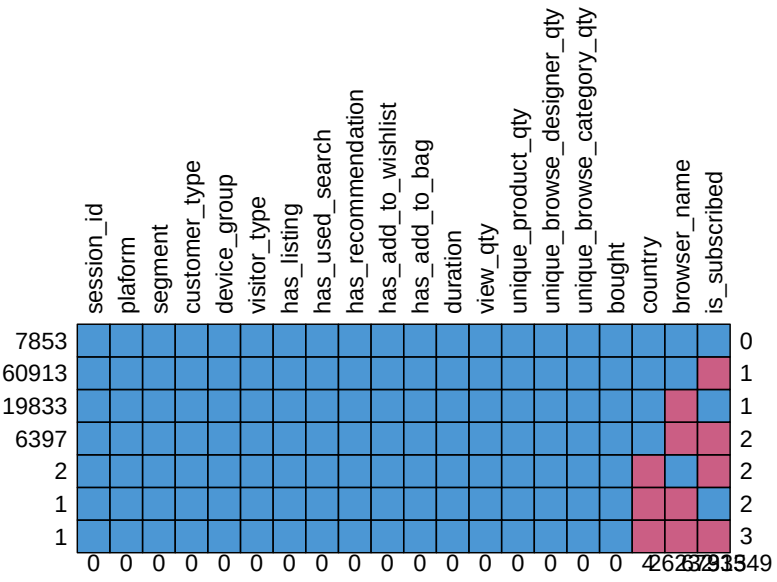
```
imputed <- bake(ratio_recipe2,new_data = NULL)
summary(imputed)
```

```
##        plaform                          segment       customer_type
##  mobile_app:26232   FFACCESS-Bronze        : 9331   customer:19471
##  website   :68768   FFACCESS-Gold          : 2405   prospect:75529
##                     FFACCESS-Platinum      : 1240
##                     FFACCESS-Private-Client: 1564
##                     FFACCESS-Silver        : 2511
##                     without_segment        :77949
##
##      device_group      visitor_type   has_listing has_used_search
##  App       :26232   new      :48592   0:44979     0:80466
##  Desktop   :16201   returning:46408   1:50021     1:14534
##  Mobile Web:52567
##
##
##
##
##  has_recommendation has_add_to_wishlist has_add_to_bag    duration
##  0:65594            0:90468             0:84191        Min.   :    0.0
##  1:29406            1: 4532             1:10809        1st Qu.:    0.0
##                                                        Median :   15.0
##                                                        Mean   :  332.5
##                                                        3rd Qu.:  223.0
##                                                        Max.   :20295.0
##
##     view_qty       unique_product_qty unique_browse_designer_qty
```

```
##  Min.   :   1.00   Min.   :  0.000    Min.   :  0.000
##  1st Qu.:   1.00   1st Qu.:  0.000    1st Qu.:  1.000
##  Median :   2.00   Median :  1.000    Median :  1.000
##  Mean   :  10.97   Mean   :  2.767    Mean   :  1.676
##  3rd Qu.:   8.00   3rd Qu.:  2.000    3rd Qu.:  1.000
##  Max.   :1151.00   Max.   :416.000    Max.   :158.000
##
##  unique_browse_category_qty is_subscribed           browser_name
##  1      :53094              No     :39612   Safari         :39038
##  0      :22498              Unknown:   44   Chrome         :33345
##  2      :11476              Yes    :55344   Instagram App  : 5076
##  3      : 3834                               Facebook App   : 3779
##  4      : 1795                               Google App     : 2567
##  5      :  929                               Samsung Browser: 2182
##  (Other): 1374                               (Other)        : 9013
##     country      bought
##  US     :15772   0:87000
##  RU     :14419   1: 8000
##  BR     : 6046
##  GB     : 5874
##  IT     : 4152
##  MX     : 3903
##  (Other):44834
```

3.3 Assess data quality & transformations to be made

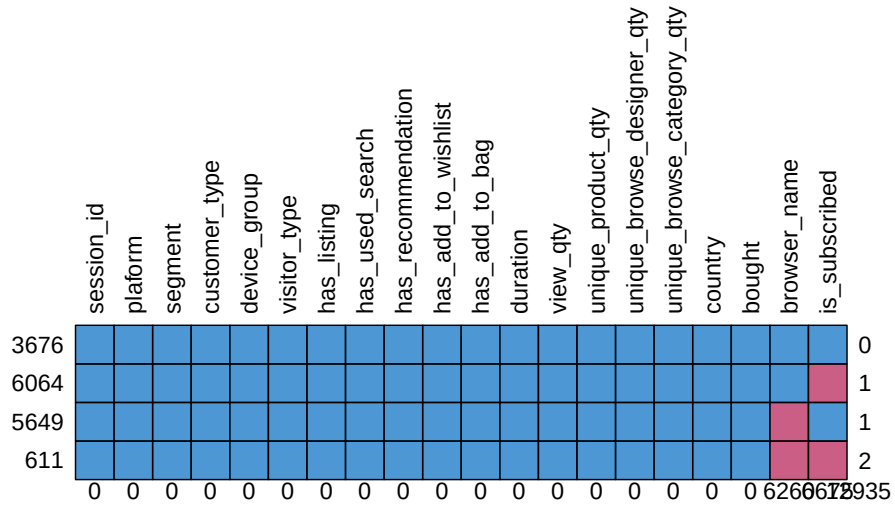
```
md.pattern(unbalanced_data, rotate.names = TRUE)
```




```
##       session_id plaform segment customer_type device_group visitor_type
## 7853           1       1       1             1            1            1
## 60913          1       1       1             1            1            1
## 19833          1       1       1             1            1            1
## 6397           1       1       1             1            1            1
## 2              1       1       1             1            1            1
## 1              1       1       1             1            1            1
## 1              1       1       1             1            1            1
##                0       0       0             0            0            0
##       has_listing has_used_search has_recommendation has_add_to_wishlist
## 7853            1                1                  1                   1
## 60913           1                1                  1                   1
## 19833           1                1                  1                   1
## 6397            1                1                  1                   1
## 2               1                1                  1                   1
## 1               1                1                  1                   1
## 1               1                1                  1                   1
##                 0                0                  0                   0
##       has_add_to_bag duration view_qty unique_product_qty
## 7853               1        1        1                  1
## 60913              1        1        1                  1
## 19833              1        1        1                  1
## 6397               1        1        1                  1
## 2                  1        1        1                  1
## 1                  1        1        1                  1
```

```
## 1                   1        1        1                   1
##                     0        0        0                   0
##       unique_browse_designer_qty unique_browse_category_qty bought country
## 7853                           1                           1      1       1
## 60913                          1                           1      1       1
## 19833                          1                           1      1       1
## 6397                           1                           1      1       1
## 2                              1                           1      1       0
## 1                              1                           1      1       0
## 1                              1                           1      1       0
##                                0                           0      0       4
##       browser_name is_subscribed
## 7853             1             1     0
## 60913            1             0     1
## 19833            0             1     1
## 6397             0             0     2
## 2                1             0     2
## 1                0             1     2
## 1                0             0     3
##              26232         67313 93549
```

```
md.pattern(balanced_data,rotate.names = TRUE)
```




```
##      session_id plaform segment customer_type device_group visitor_type
```

```
## 3676          1       1       1             1             1             1
## 6064          1       1       1             1             1             1
## 5649          1       1       1             1             1             1
## 611           1       1       1             1             1             1
##               0       0       0             0             0             0
##      has_listing has_used_search has_recommendation has_add_to_wishlist
## 3676           1                1                  1                    1
## 6064           1                1                  1                    1
## 5649           1                1                  1                    1
## 611            1                1                  1                    1
##                0                0                  0                    0
##      has_add_to_bag duration view_qty unique_product_qty
## 3676              1        1        1                  1
## 6064              1        1        1                  1
## 5649              1        1        1                  1
## 611               1        1        1                  1
##                   0        0        0                  0
##      unique_browse_designer_qty unique_browse_category_qty country bought
## 3676                          1                           1       1      1
## 6064                          1                           1       1      1
## 5649                          1                           1       1      1
## 611                           1                           1       1      1
##                               0                           0       0      0
##      browser_name is_subscribed
## 3676            1             1     0
## 6064            1             0     1
## 5649            0             1     1
## 611             0             0     2
##              6260          6675 12935
```

3.4 Summary of findings

- Device group and platform have a strong relationship and should be taken into consideration during transformations.
- Features conserning journey can have strong intereactions between then which should be taken into consideration during the model
- The graphical analysis already provides a important insight given the business objectives. From the unbalanced data available we can conclude that around of 46% shopping carts are lost on that session. That raises a question of how are this recovered (example on a next session) or if this means that all this sales are lost right at the end of the sales funnel.
- Log transform duration
- Log transform view_qty and remove below or after tukeys value
- Log transform unique_qty remove below or after tukeys value

- Log transform nique browse designer quantity remove below or after tukeys value
- remove rows with na country