# R Notebook

## leo datos

```
data<-read.csv("entrena.csv",stringsAsFactors = F)
```

## Creo las 3 bases

```
knitr::opts_chunk$set( message = FALSE, results='hide')
idx <- sample(seq(1, 3), size = nrow(data), replace = TRUE, prob = c(.7, 2, .1))
train <- data[idx == 1,]
test <- data[idx == 2,]
cal <- data[idx == 3,]
```

## summary

```
knitr::opts_chunk$set(echo = TRUE, message = FALSE, results = 'show')
summary(train)
```

```
##     hotel           is_canceled          lead_time      arrival_date_year
##  Length:22972       Length:22972       Min.   :  0.00   Min.   :2015
##  Class :character   Class :character   1st Qu.: 14.00   1st Qu.:2016
##  Mode  :character   Mode  :character   Median : 57.00   Median :2016
##                                        Mean   : 96.12   Mean   :2016
##                                        3rd Qu.:144.00   3rd Qu.:2016
##                                        Max.   :737.00   Max.   :2017
##
##  arrival_date_month arrival_date_week_number arrival_date_day_of_month
##  Length:22972       Min.   : 1.00            Min.   : 1.00
##  Class :character   1st Qu.:13.00            1st Qu.: 8.00
##  Mode  :character   Median :30.00            Median :16.00
##                     Mean   :28.13            Mean   :15.79
##                     3rd Qu.:41.00            3rd Qu.:23.00
##                     Max.   :53.00            Max.   :31.00
##
##  stays_in_weekend_nights stays_in_week_nights     adults
##  Min.   : 0.0000         Min.   : 0.000       Min.   : 0.000
##  1st Qu.: 0.0000         1st Qu.: 1.000       1st Qu.: 2.000
##  Median : 1.0000         Median : 2.000       Median : 2.000
##  Mean   : 0.8931         Mean   : 2.459       Mean   : 1.835
##  3rd Qu.: 2.0000         3rd Qu.: 3.000       3rd Qu.: 2.000
##  Max.   :19.0000         Max.   :50.000       Max.   :55.000
##
##     children           babies             meal             country
##  Min.   :0.00000   Min.   :0.000000   Length:22972       Length:22972
##  1st Qu.:0.00000   1st Qu.:0.000000   Class :character   Class :character
##  Median :0.00000   Median :0.000000   Mode  :character   Mode  :character
##  Mean   :0.08755   Mean   :0.007487
##  3rd Qu.:0.00000   3rd Qu.:0.000000
##  Max.   :3.00000   Max.   :2.000000
##  NA's   :1
##  market_segment     distribution_channel is_repeated_guest
##  Length:22972       Length:22972         Min.   :0.00000
##  Class :character   Class :character     1st Qu.:0.00000
##  Mode  :character   Mode  :character     Median :0.00000
##                                          Mean   :0.03334
##                                          3rd Qu.:0.00000
##                                          Max.   :1.00000
##
##  previous_cancellations previous_bookings_not_canceled reserved_room_type
##  Min.   : 0.0000        Min.   : 0.0000                Length:22972
##  1st Qu.: 0.0000        1st Qu.: 0.0000                Class :character
##  Median : 0.0000        Median : 0.0000                Mode  :character
##  Mean   : 0.1133        Mean   : 0.1371
##  3rd Qu.: 0.0000        3rd Qu.: 0.0000
##  Max.   :26.0000        Max.   :61.0000
##
##  assigned_room_type booking_changes  deposit_type          agent
##  Length:22972       Min.   : 0.000   Length:22972       Length:22972
##  Class :character   1st Qu.: 0.000   Class :character   Class :character
##  Mode  :character   Median : 0.000   Mode  :character   Mode  :character
##                     Mean   : 0.214
##                     3rd Qu.: 0.000
##                     Max.   :21.000
##
##    company          days_in_waiting_list customer_type           adr
```

```
##  Length:22972       Min.   :  0.000      Length:22972       Min.   :  0.00
##  Class :character   1st Qu.:  0.000      Class :character   1st Qu.: 64.00
##  Mode  :character   Median :  0.000      Mode  :character   Median : 87.00
##                     Mean   :  3.072                         Mean   : 92.89
##                     3rd Qu.:  0.000                         3rd Qu.:114.75
##                     Max.   :391.000                         Max.   :508.00
##
##  required_car_parking_spaces total_of_special_requests
##  Min.   :0.00000             Min.   :0.0000
##  1st Qu.:0.00000             1st Qu.:0.0000
##  Median :0.00000             Median :0.0000
##  Mean   :0.06599             Mean   :0.5257
##  3rd Qu.:0.00000             3rd Qu.:1.0000
##  Max.   :2.00000             Max.   :5.0000
##
```

# proporcion de cancelados

```
a <- table(train$is_canceled)
prop <- prop.table(a)
prop
```

```
##
##    cancelado no_cancelado
##    0.3618753    0.6381247
```

```
sub_cancelados <- subset(train, is_canceled == "cancelado")
```

#funcion para graficar las variables originales

```
analisis_cancelaciones <- function(datos,nombre) {
cancel <- table(datos)
par(cex=0.5) #control size of labels
g_1 <- barplot(cancel, main = nombre)
  return(g_1)
}
```

#funcion para graficar las variables ordenadas

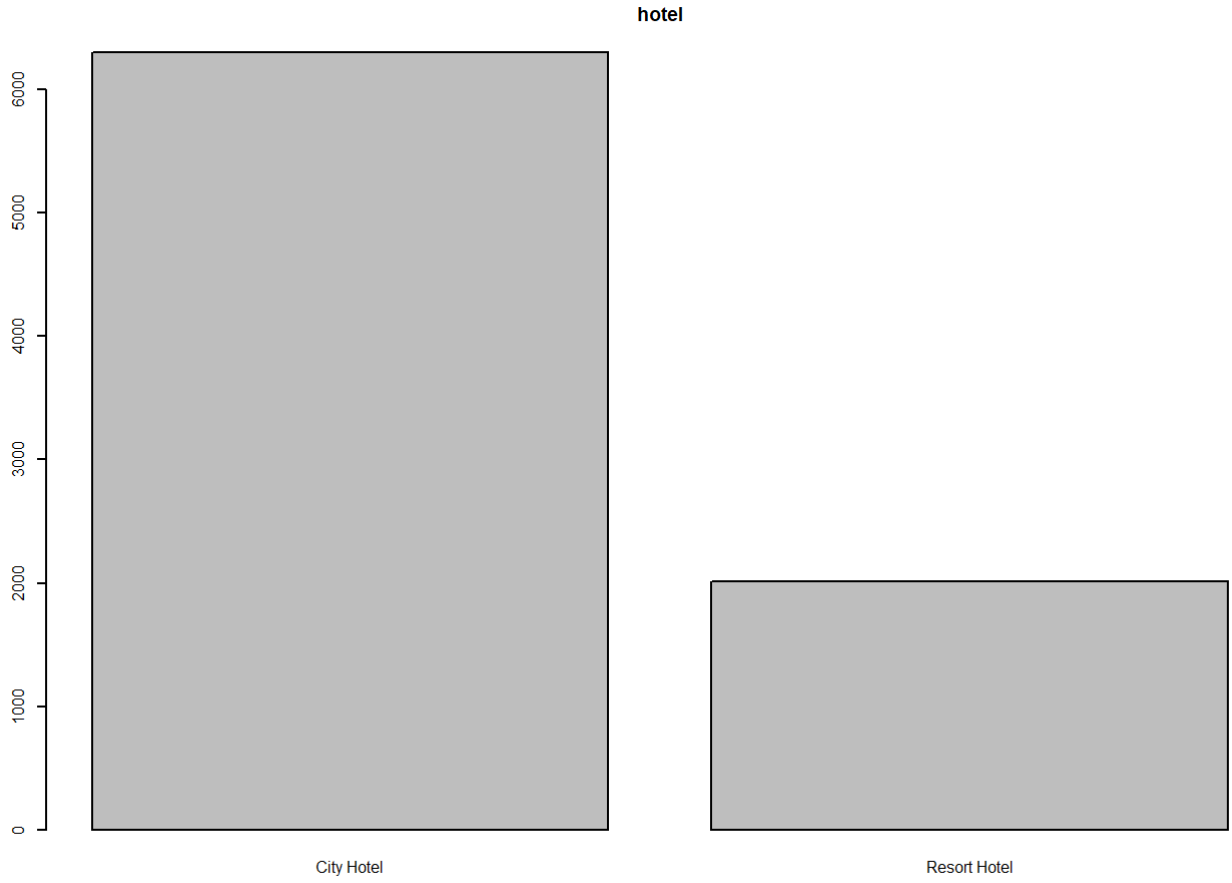
```
analisis_cancelaciones_ord <- function(datos,nombre) {
x <- table(datos)
cancel <- x[order(x,decreasing = TRUE)]
par(cex=0.5) #control size of labels
g_1 <- barplot(cancel, main = nombre)
  return(g_1)
}
```

# analisis de todas las variables con el subconjunto de canceladas

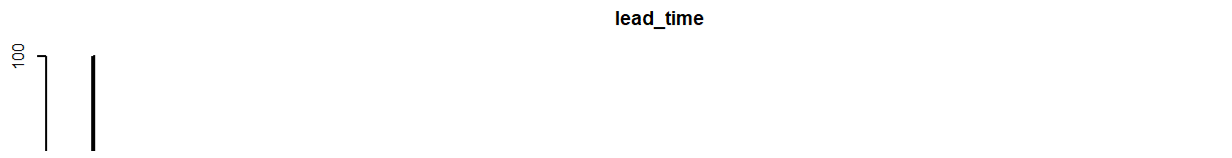
```
ncol(sub_cancelados)
```

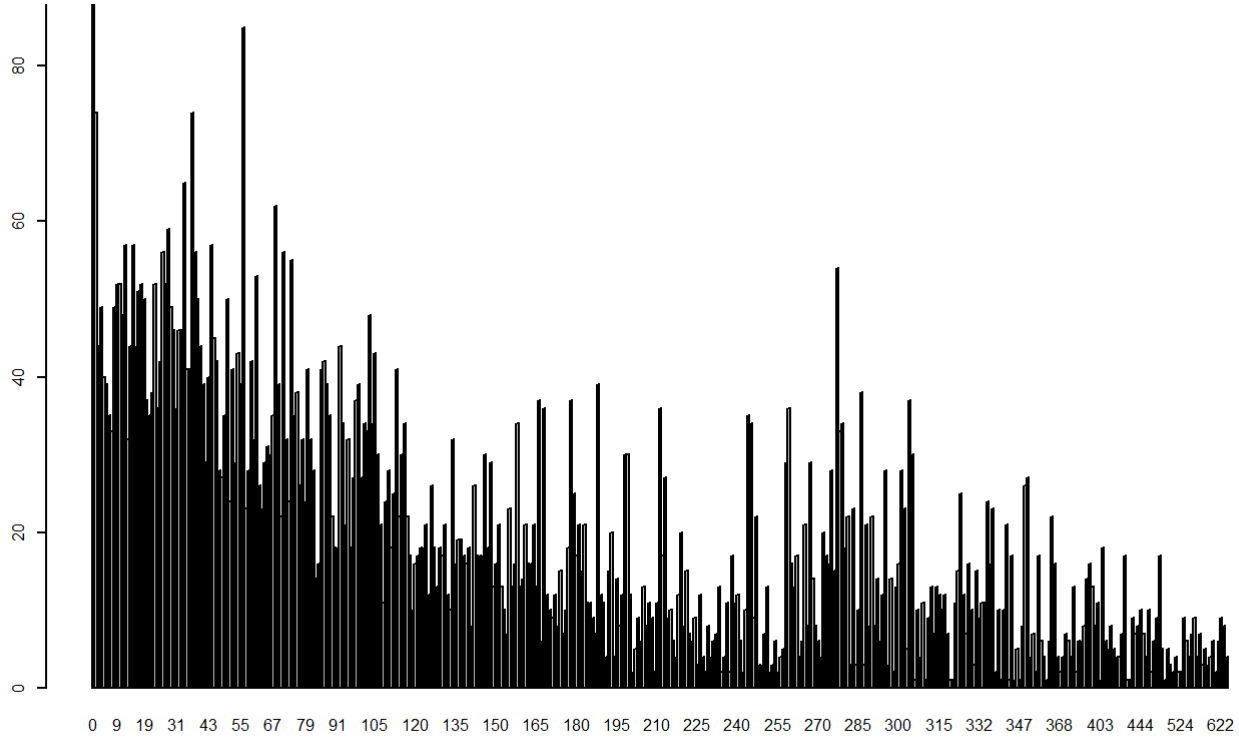
```
## [1] 30
```

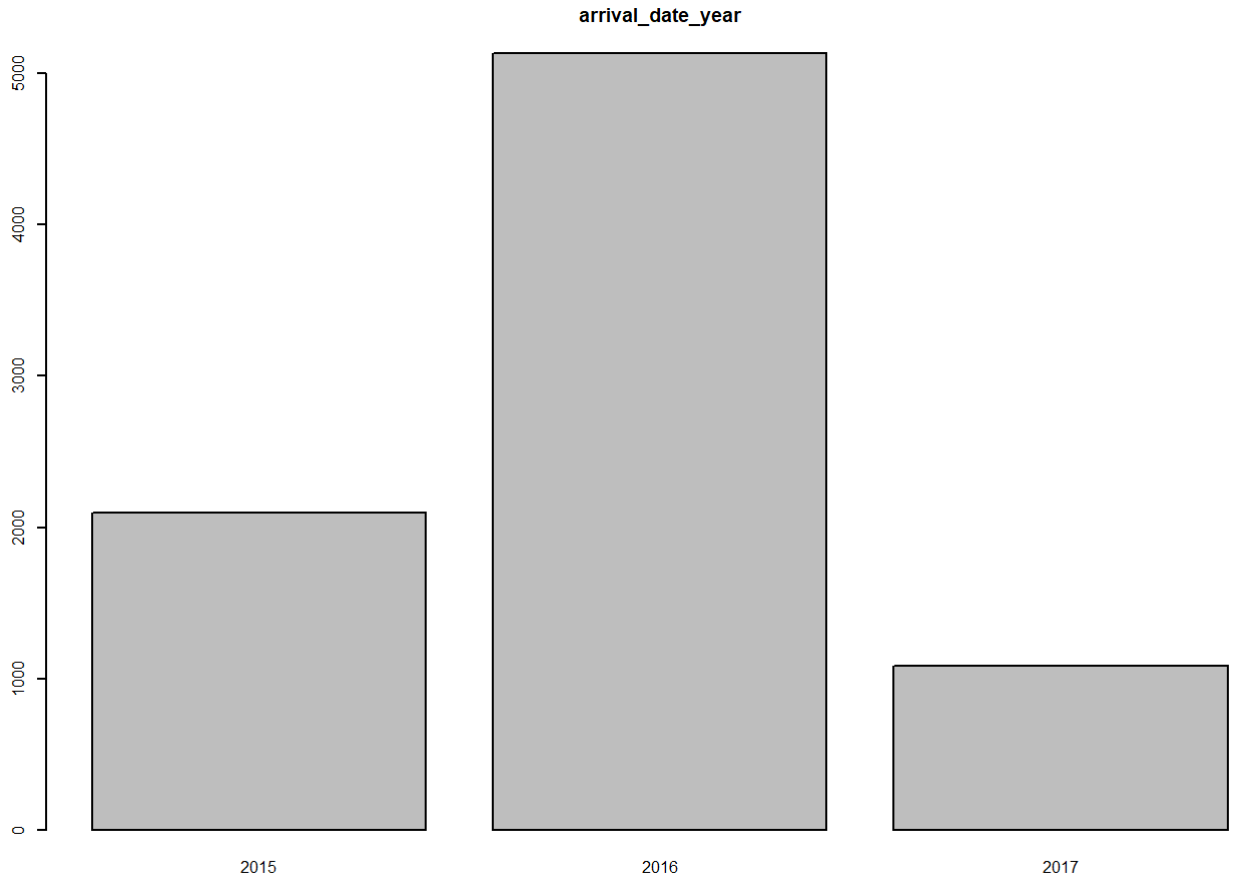
```
nombres <- names(sub_cancelados)
for (i in c(1:ncol(sub_cancelados))){
analisis_cancelaciones(sub_cancelados[,i],nombres[i])
}
```

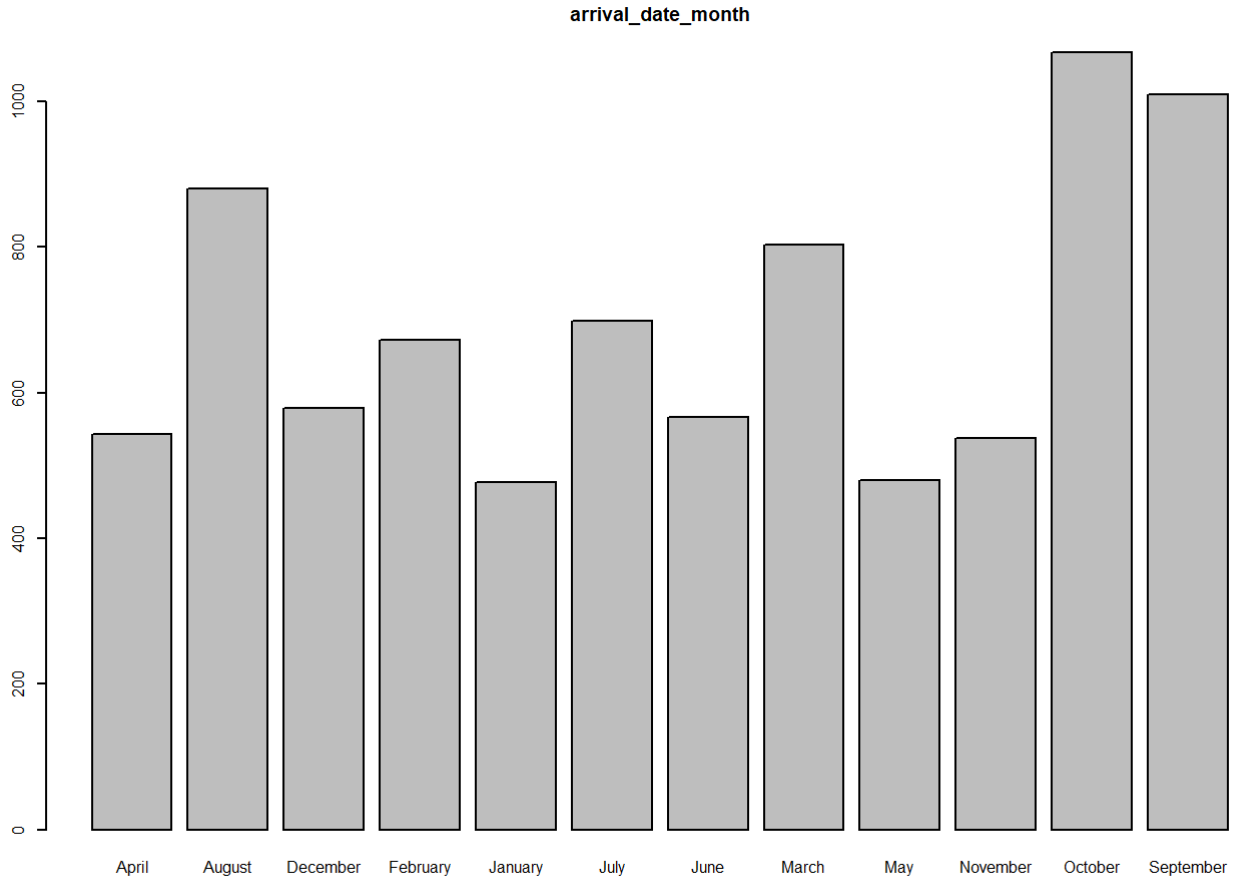

hotel
0
1000
2000
3000
4000
5000
6000
City Hotel
Resort Hotel

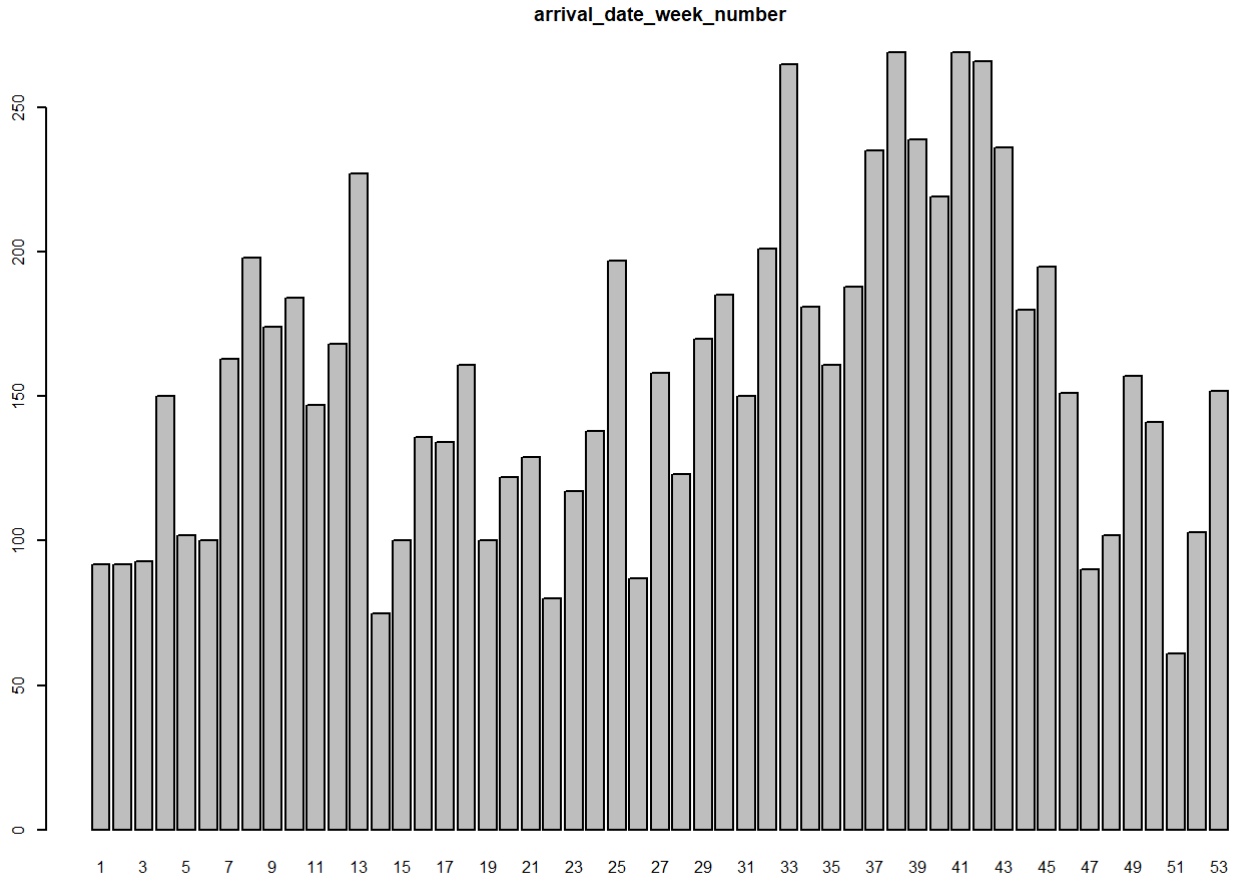

is_canceled
0
2000
4000
6000
8000
cancelado

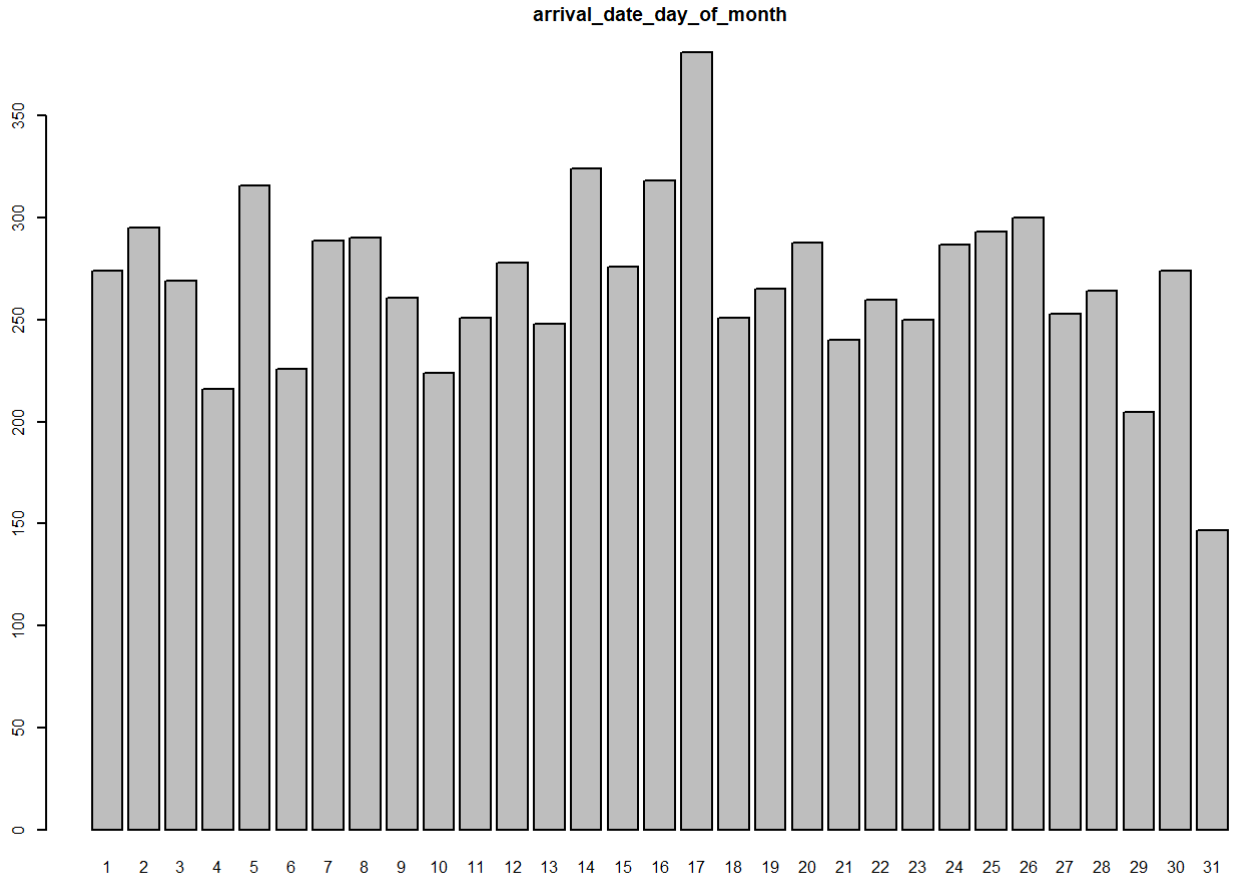

lead_time
100

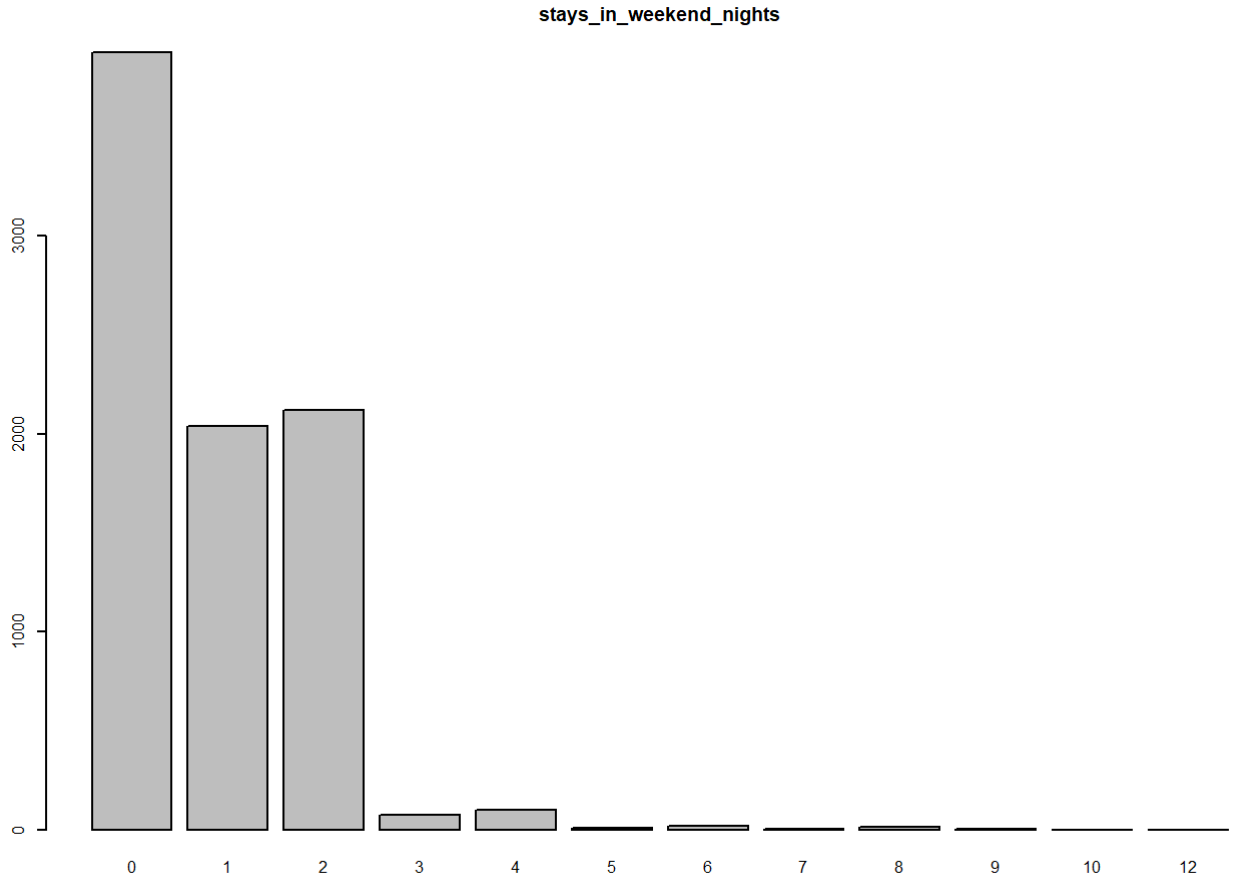

0
20
40
60
80
0 9 19 31 43 55 67 79 91 105 120 135 150 165 180 195 210 225 240 255 270 285 300 315 332 347 368 403 444 524 622

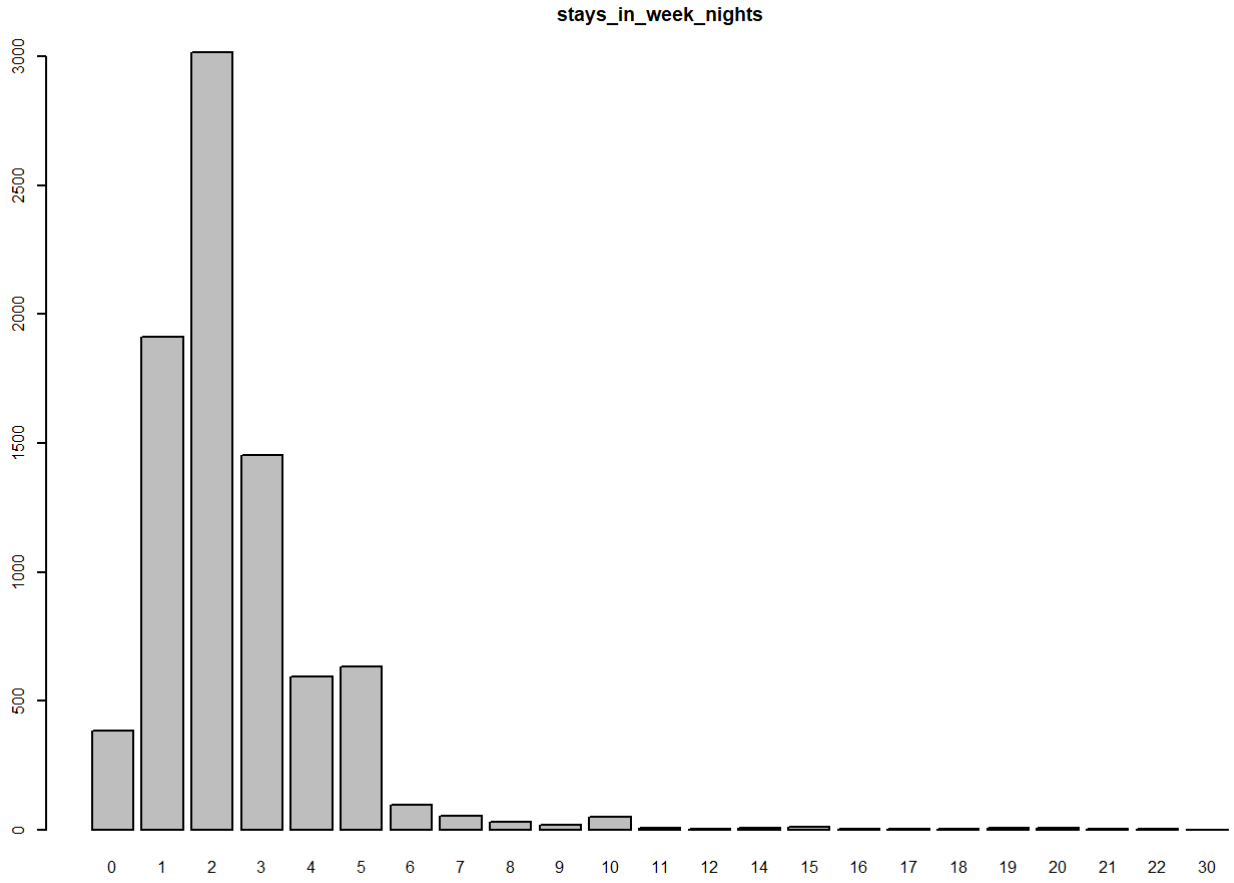

arrival_date_year
0
1000
2000
3000
4000
5000
2015
2016
2017

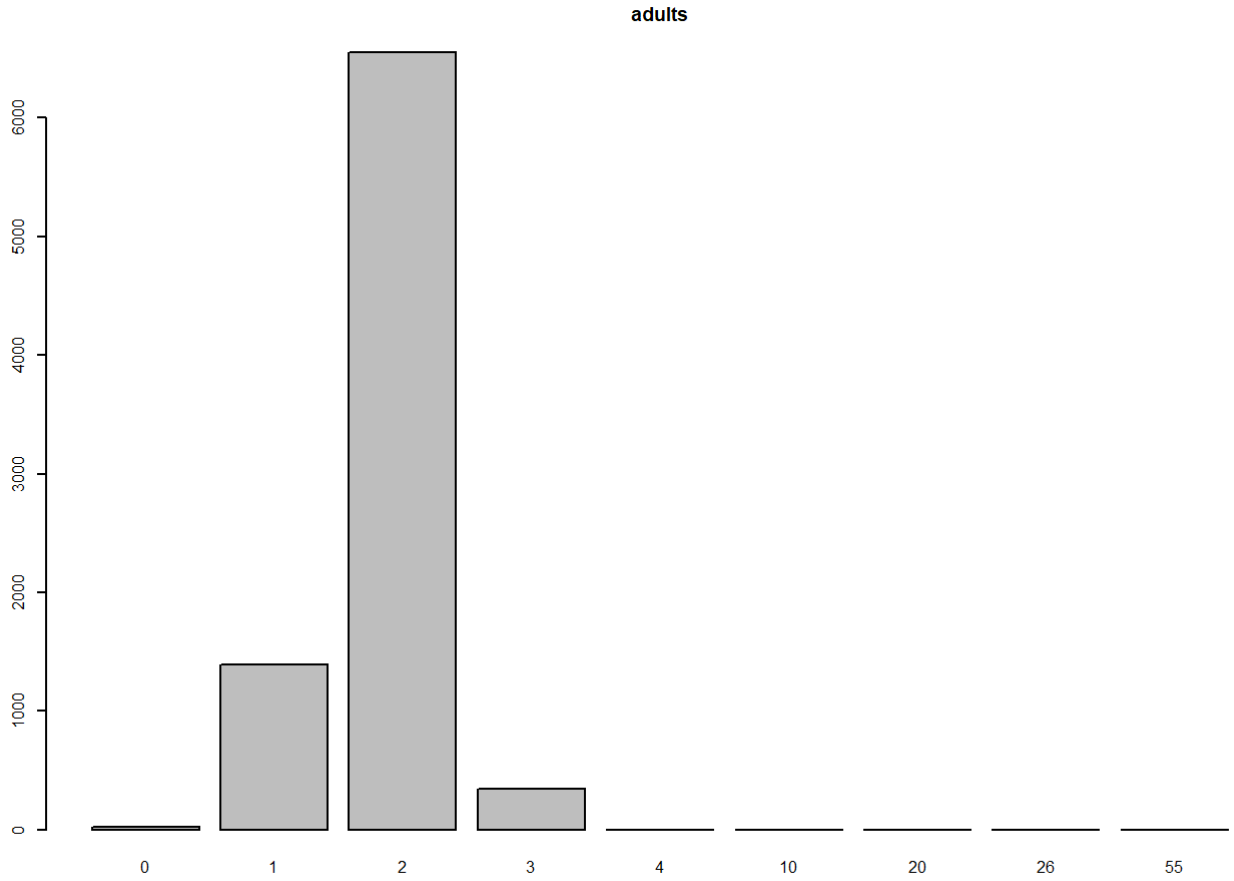

arrival_date_month
0
200
400
600
800
1000
April
August
December
February
January
July
June
March
May
November
October
September

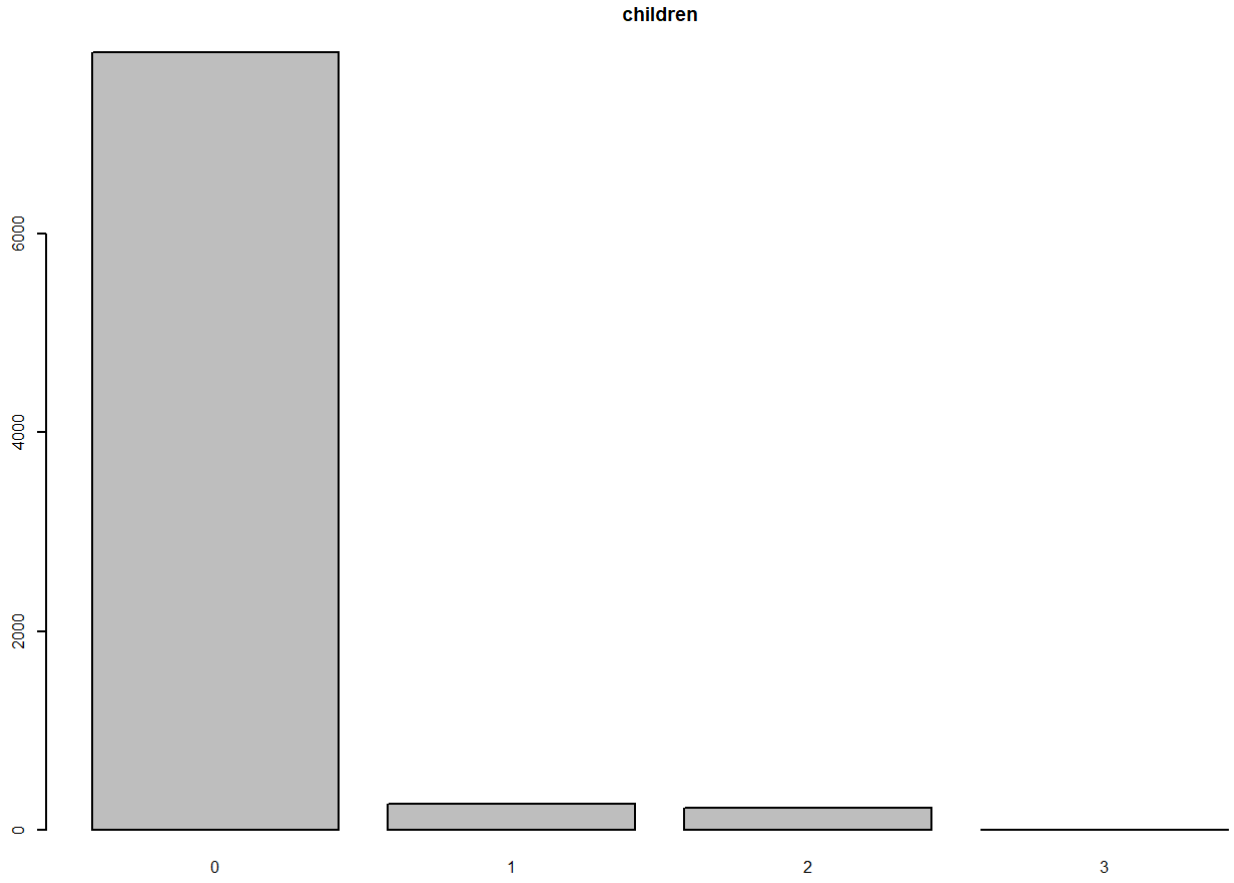

arrival_date_week_number
0
50
100
150
200
250
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
51
53

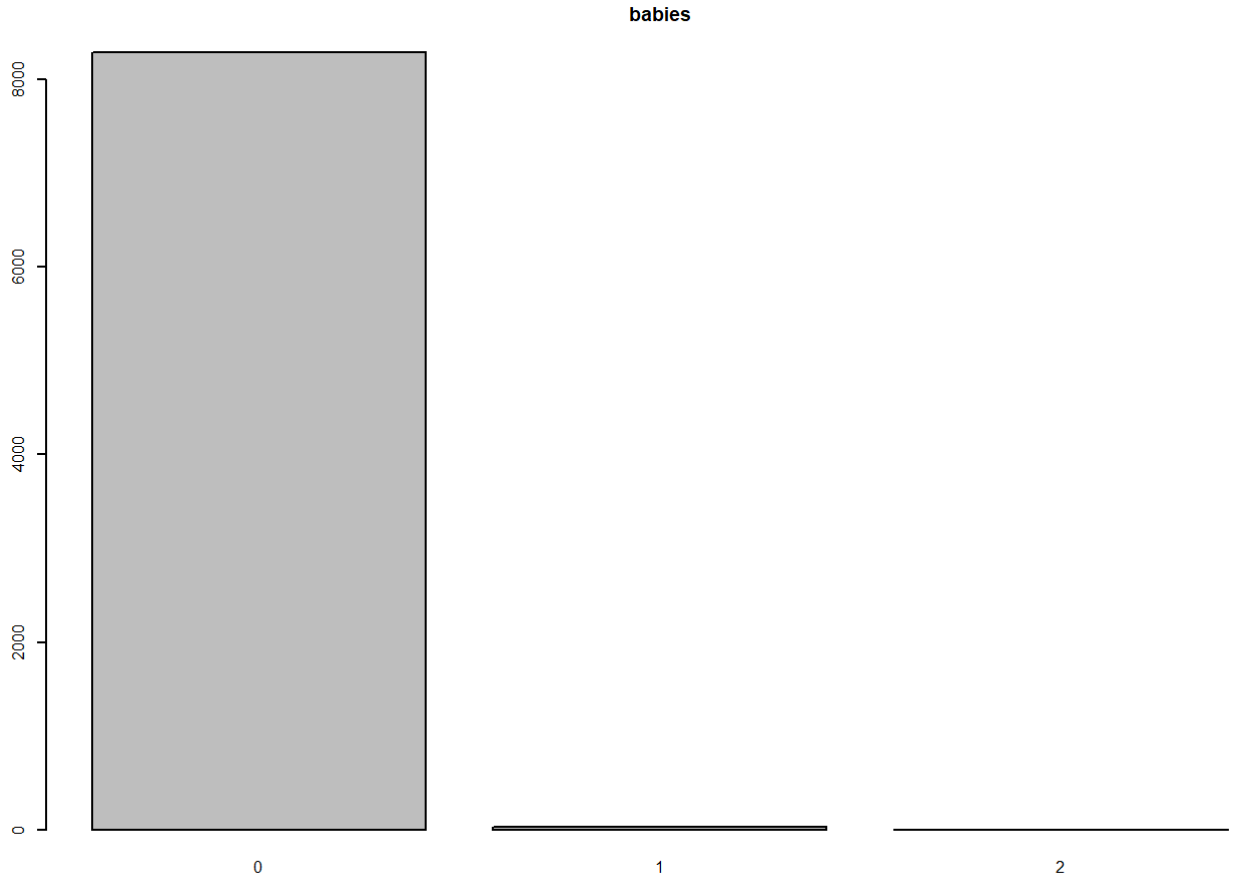

arrival_date_day_of_month
0
50
100
150
200
250
300
350
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

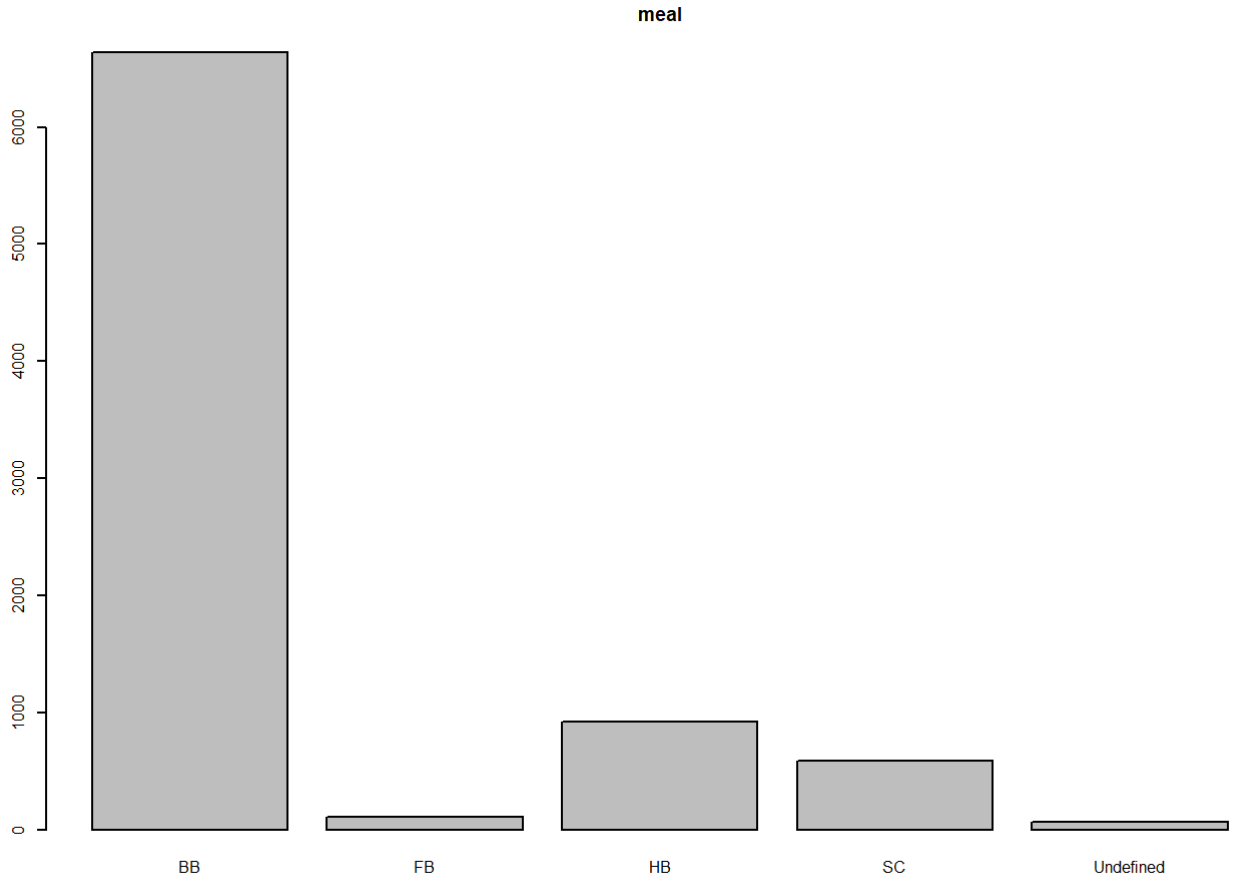

stays_in_weekend_nights
0
1000
2000
3000
0 1 2 3 4 5 6 7 8 9 10 12

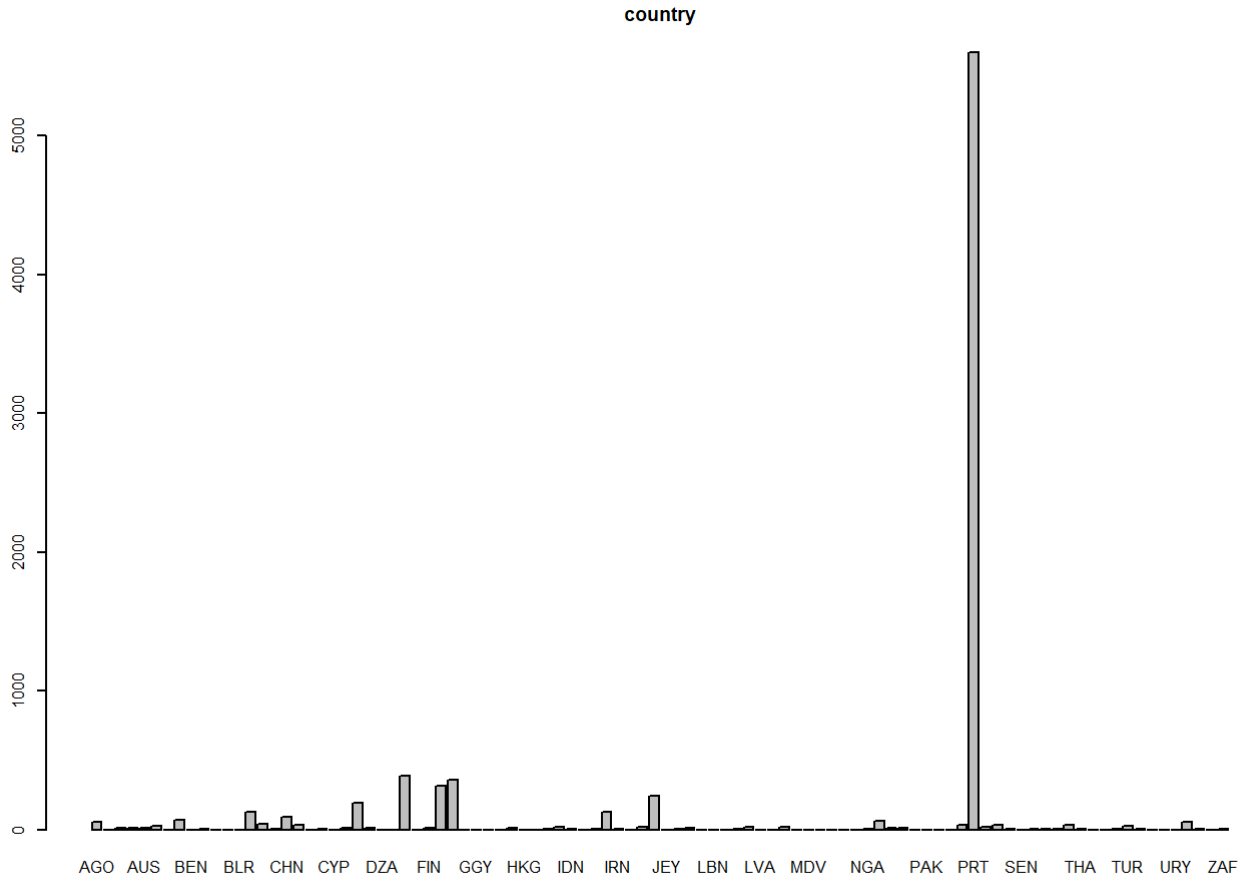

stays_in_week_nights
0
500
1000
1500
2000
2500
3000
0 1 2 3 4 5 6 7 8 9 10 11 12 14 15 16 17 18 19 20 21 22 30

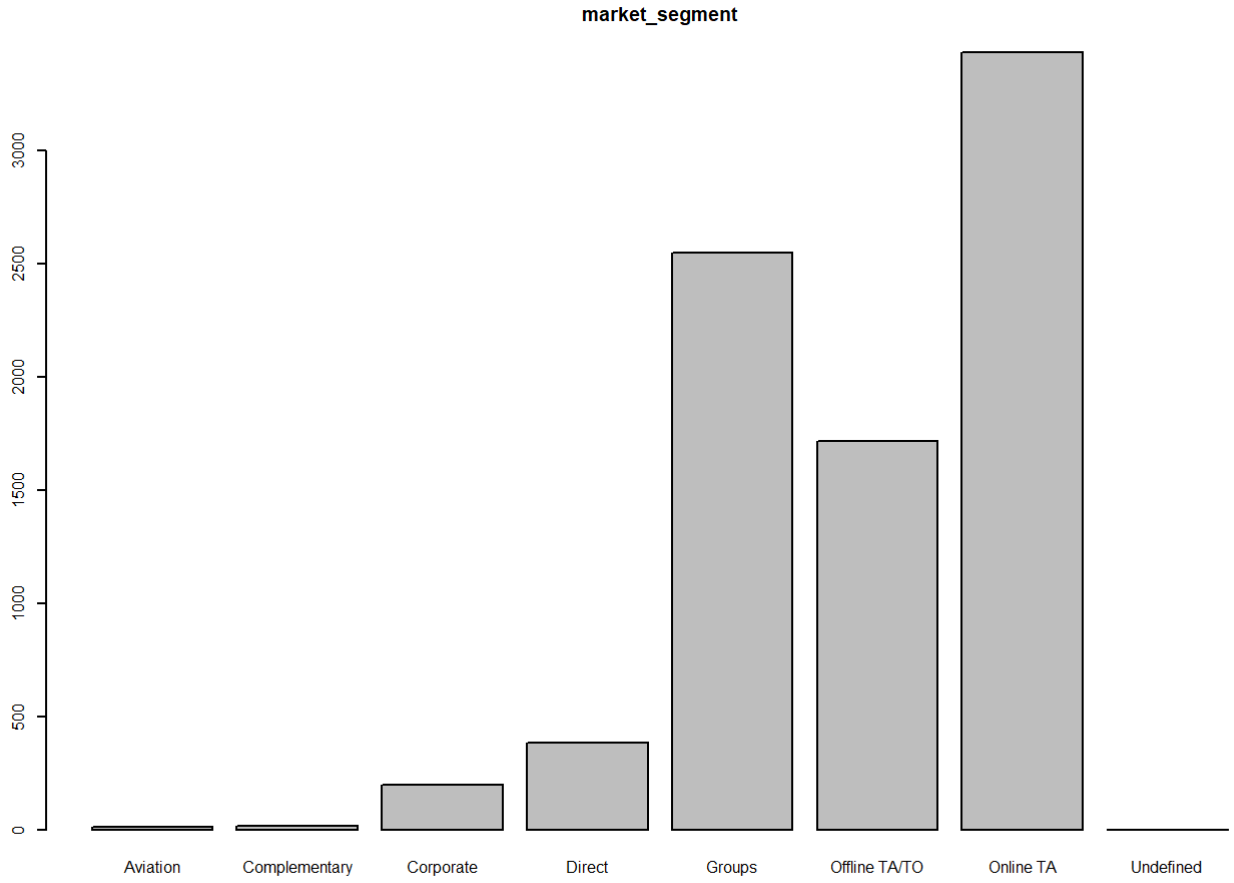

adults
0
1000
2000
3000
4000
5000
6000
0 1 2 3 4 10 20 26 55

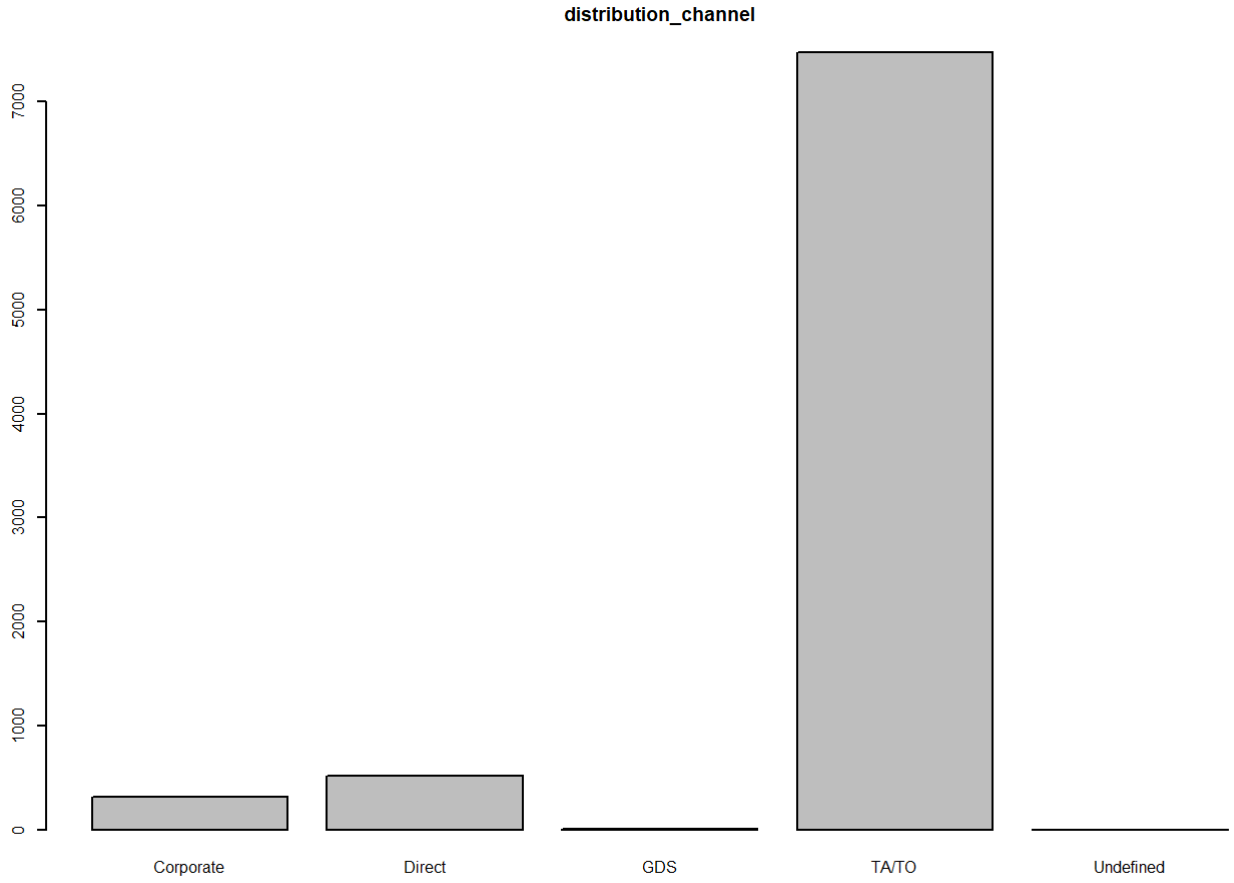

children
6000
4000
2000
0
0
1
2
3

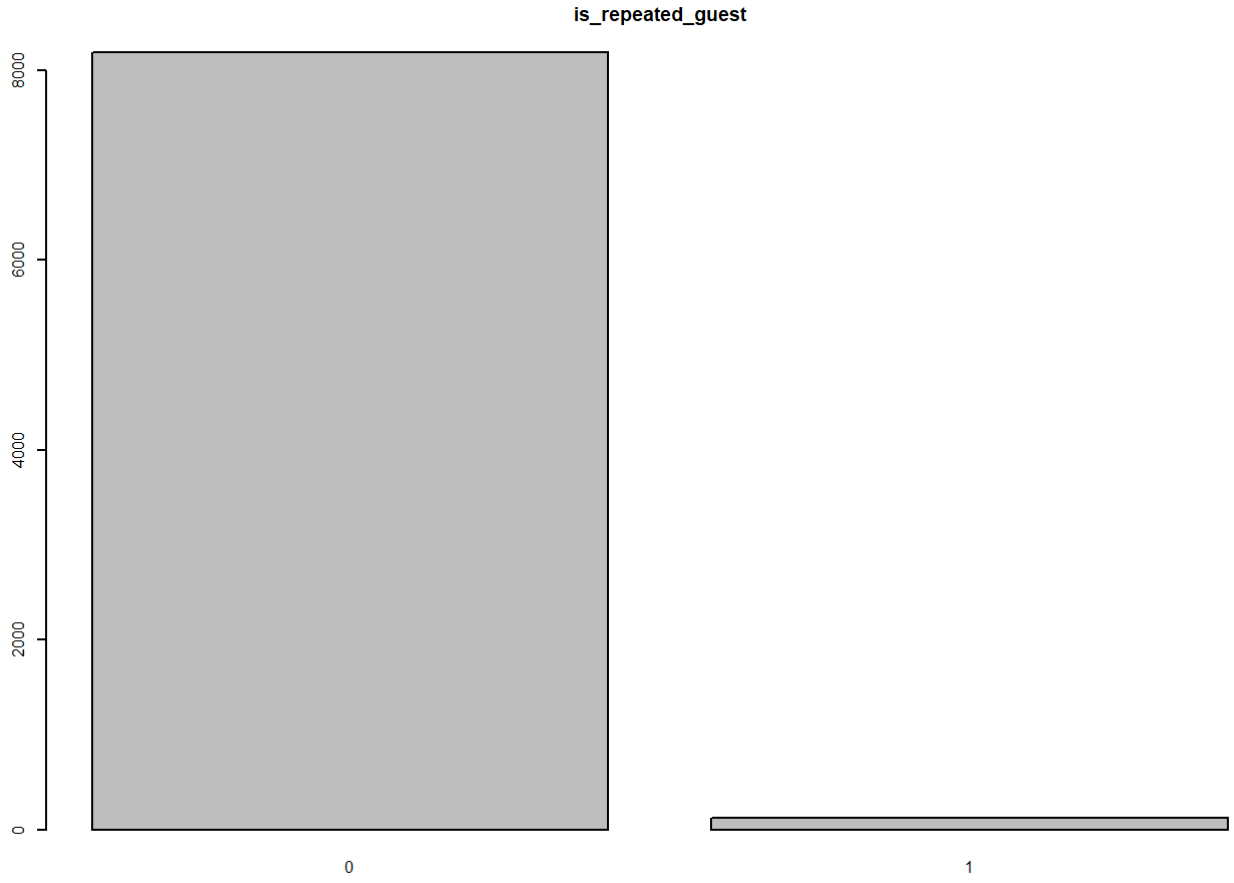

babies
8000
6000
4000
2000
0
0
1
2

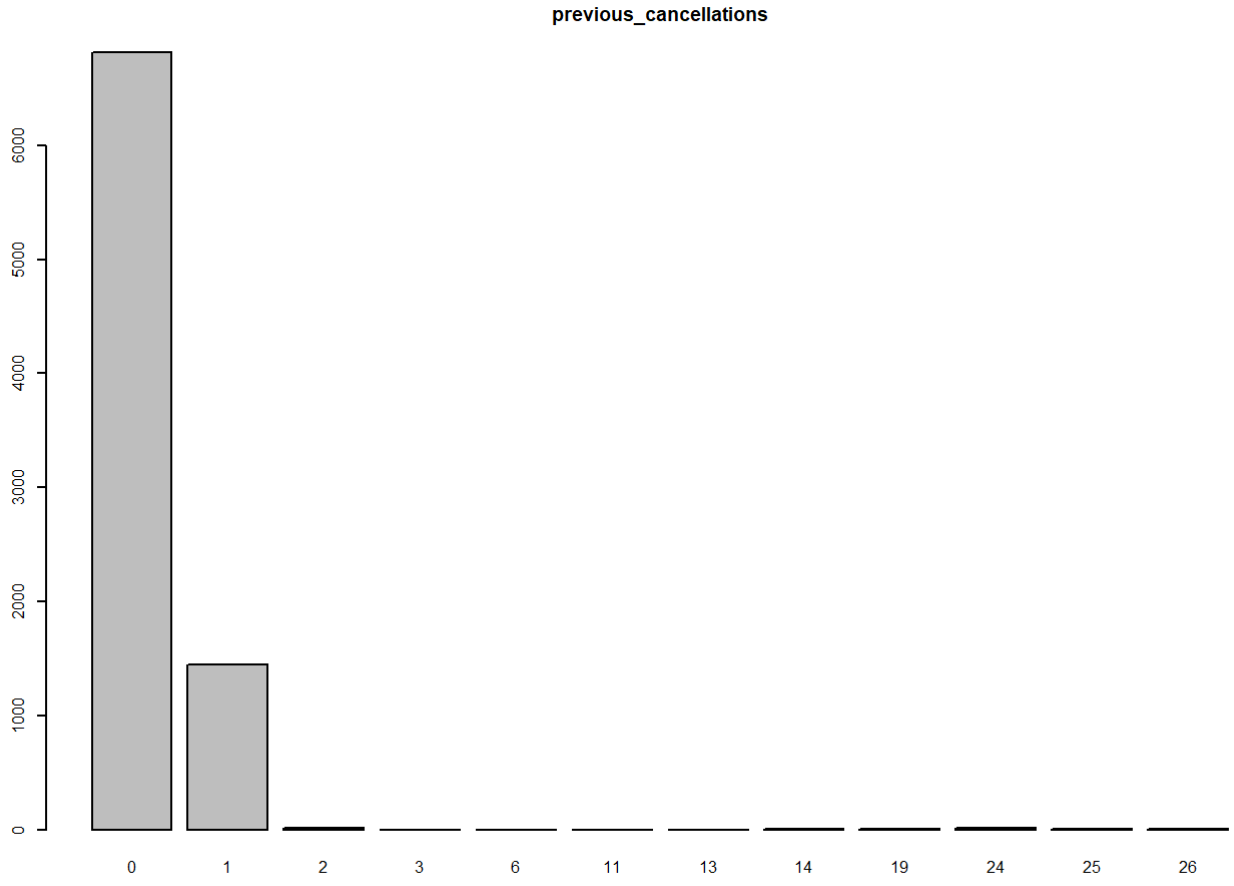

meal
0
1000
2000
3000
4000
5000
6000
BB
FB
HB
SC
Undefined

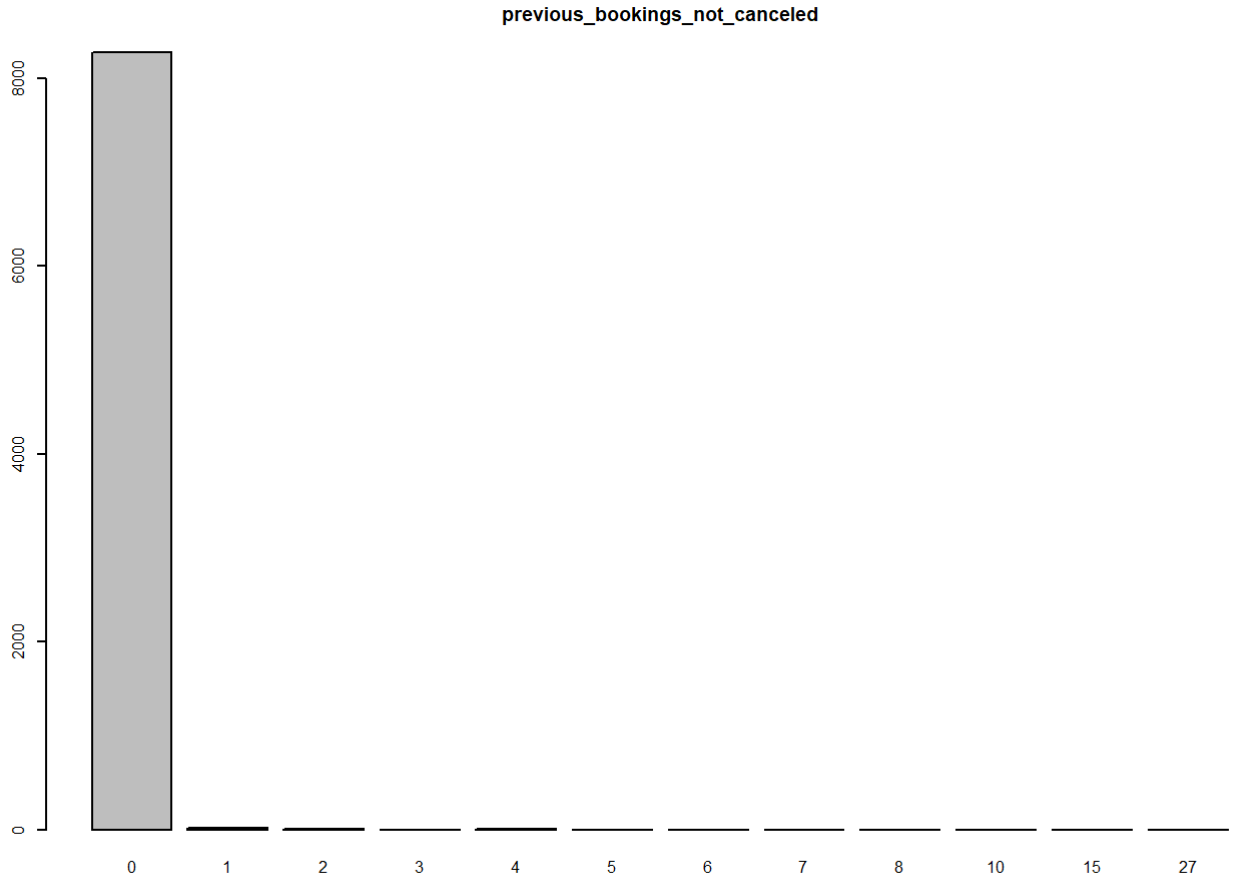

country
0
1000
2000
3000
4000
5000
AGO
AUS
BEN
BLR
CHN
CYP
DZA
FIN
GGY
HKG
IDN
IRN
JEY
LBN
LVA
MDV
NGA
PAK
PRT
SEN
THA
TUR
URY
ZAF

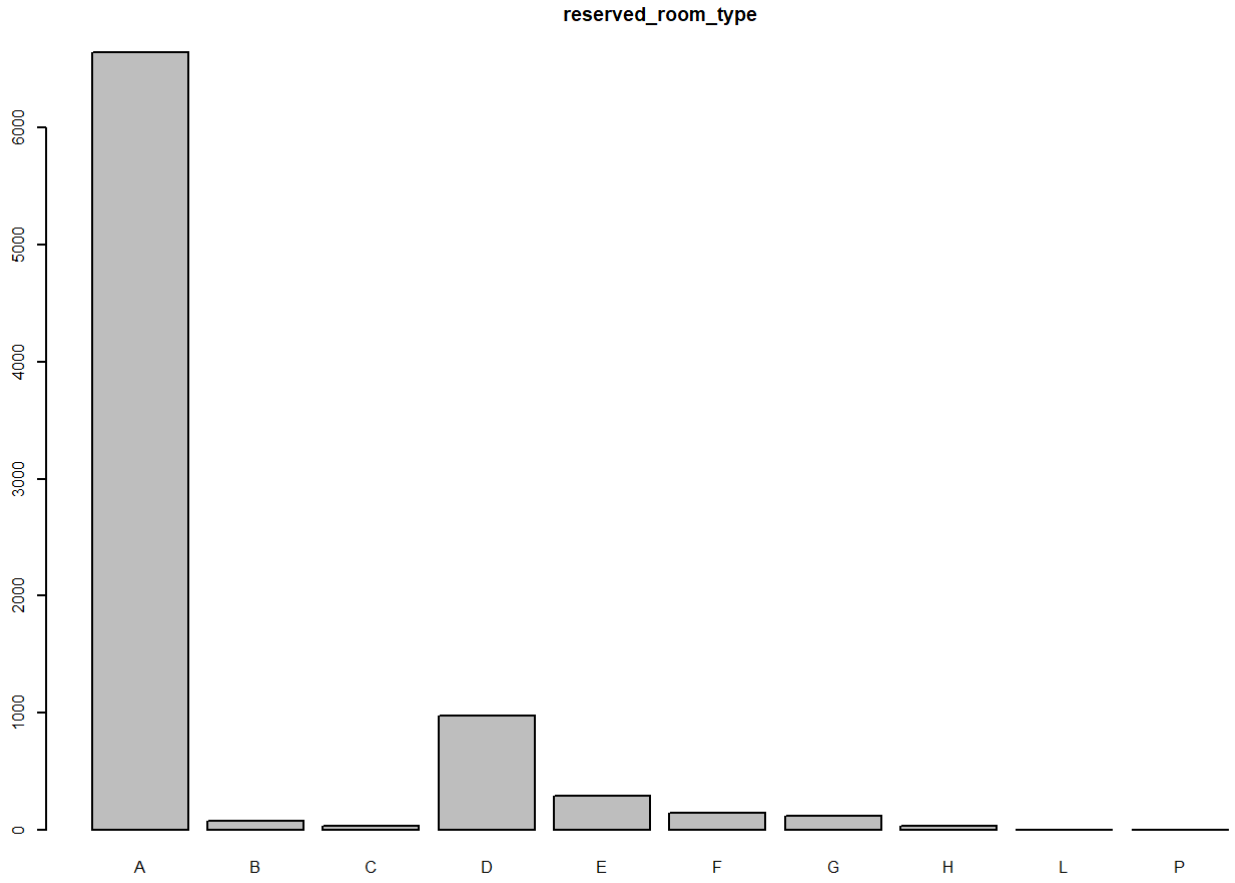

market_segment
0
500
1000
1500
2000
2500
3000
Aviation
Complementary
Corporate
Direct
Groups
Offline TA/TO
Online TA
Undefined

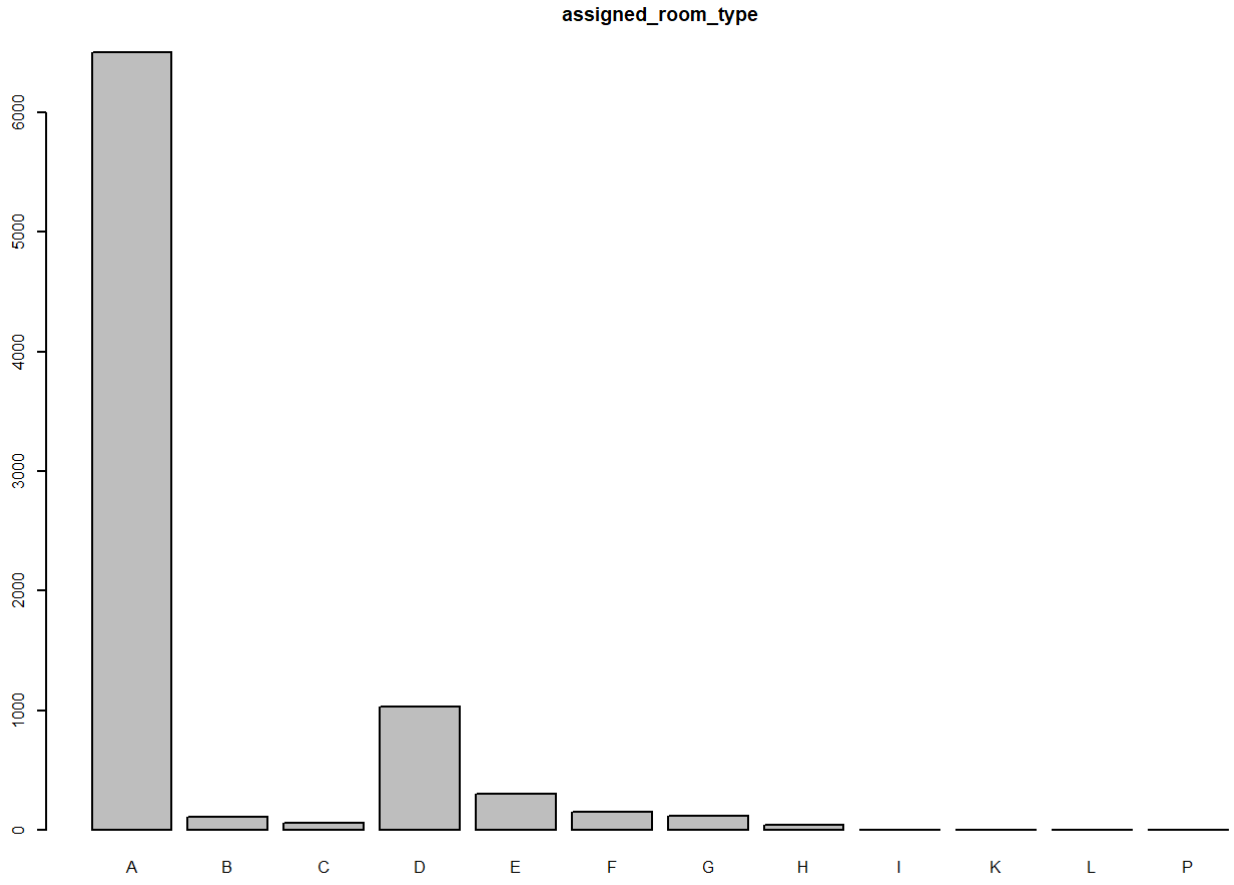

distribution_channel
0
1000
2000
3000
4000
5000
6000
7000
Corporate
Direct
GDS
TA/TO
Undefined

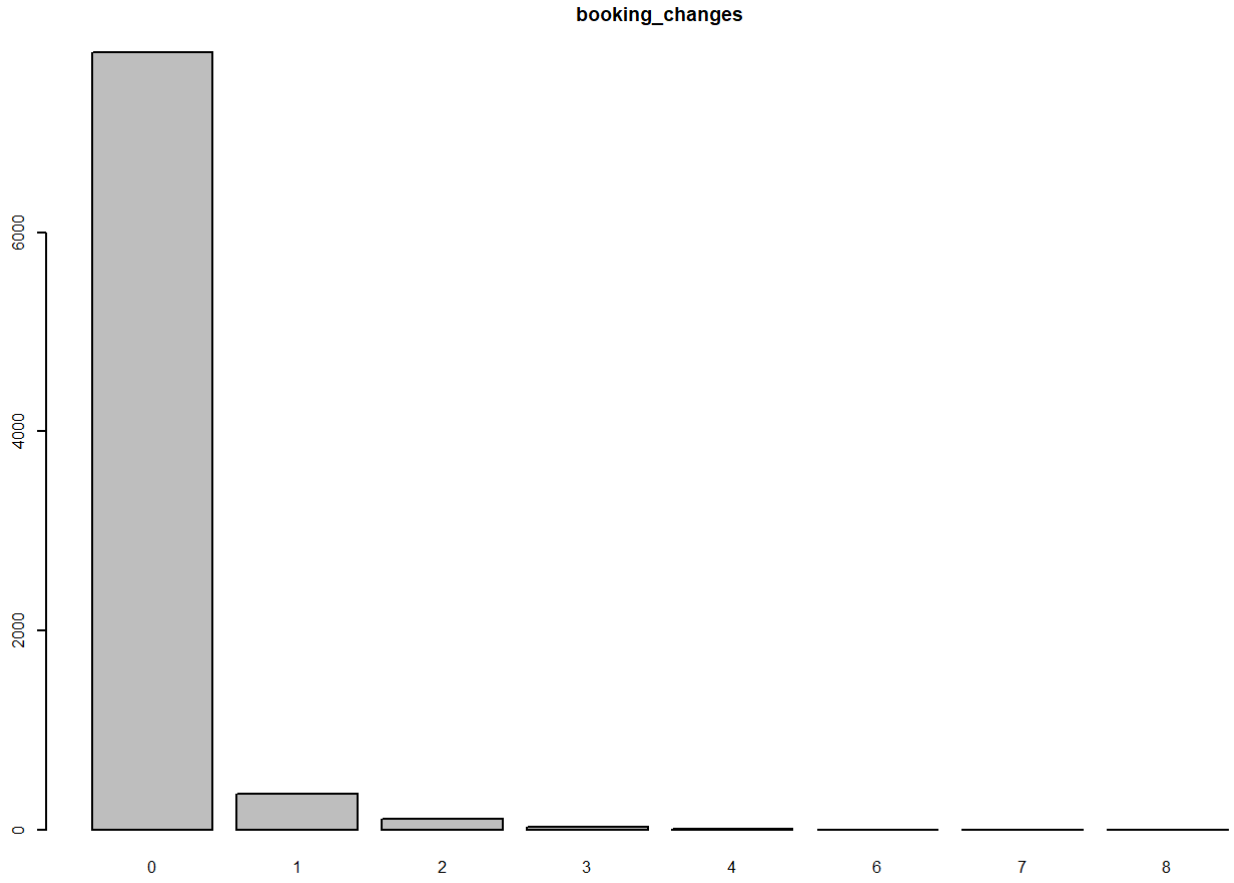

is_repeated_guest
8000
6000
4000
2000
0
0
1

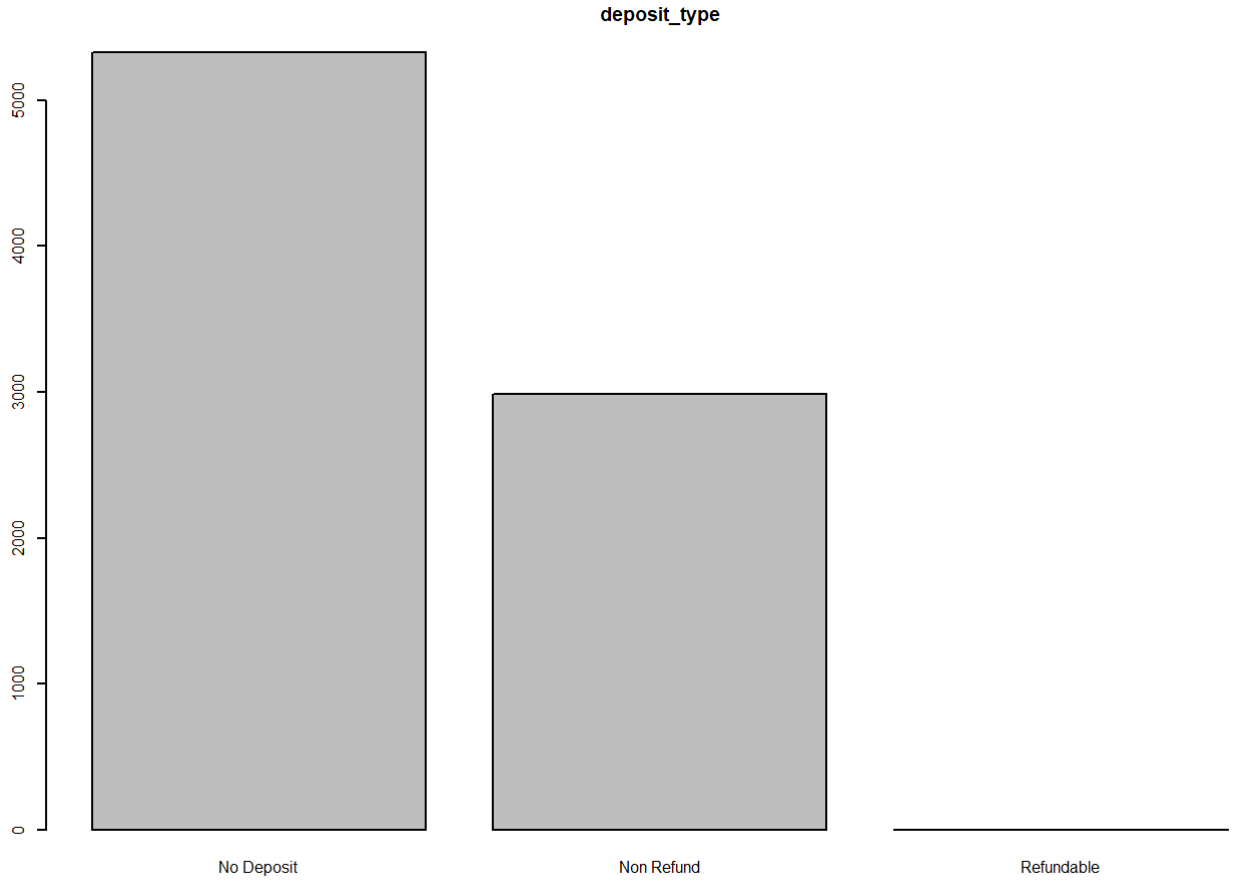

previous_cancellations
6000
5000
4000
3000
2000
1000
0
0
1
2
3
6
11
13
14
19
24
25
26

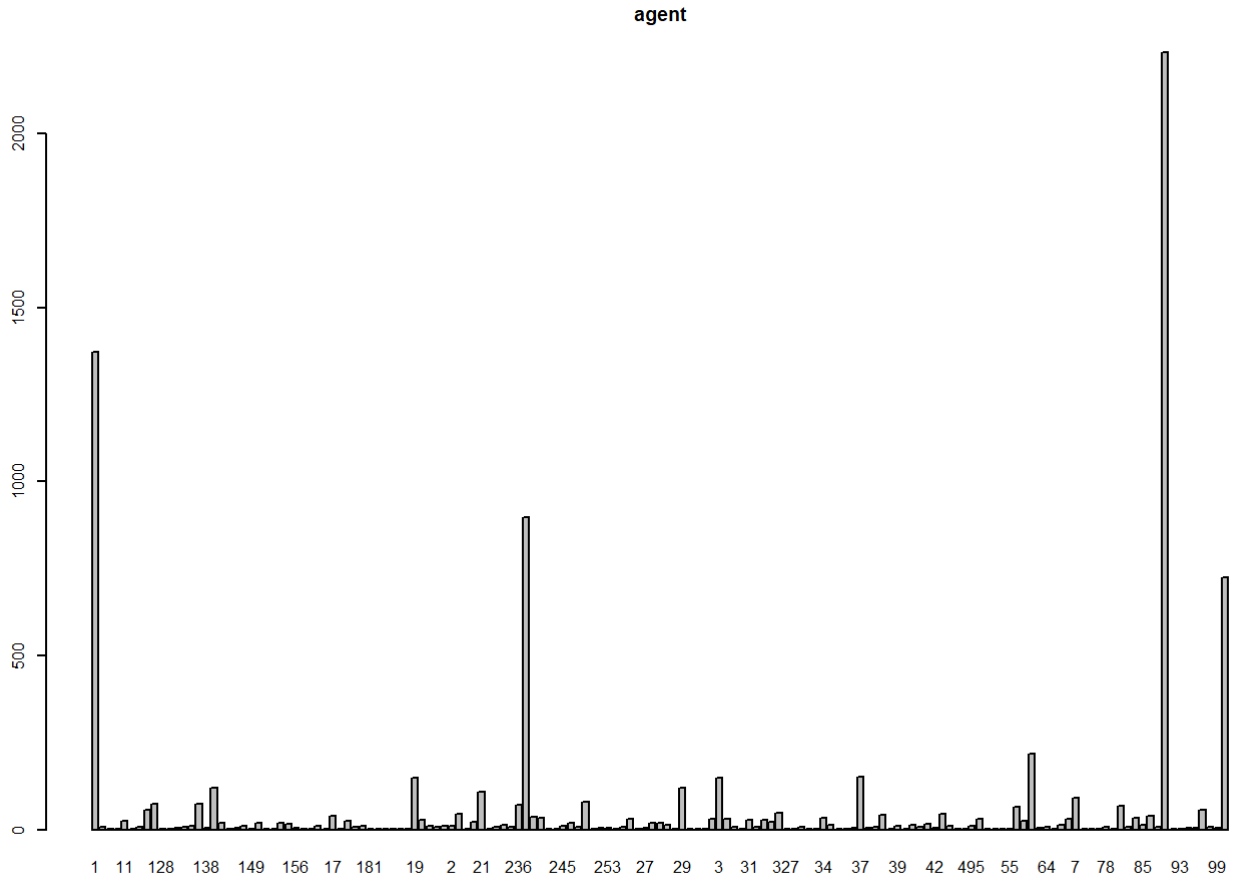

previous_bookings_not_canceled
0
2000
4000
6000
8000
0
1
2
3
4
5
6
7
8
10
15
27

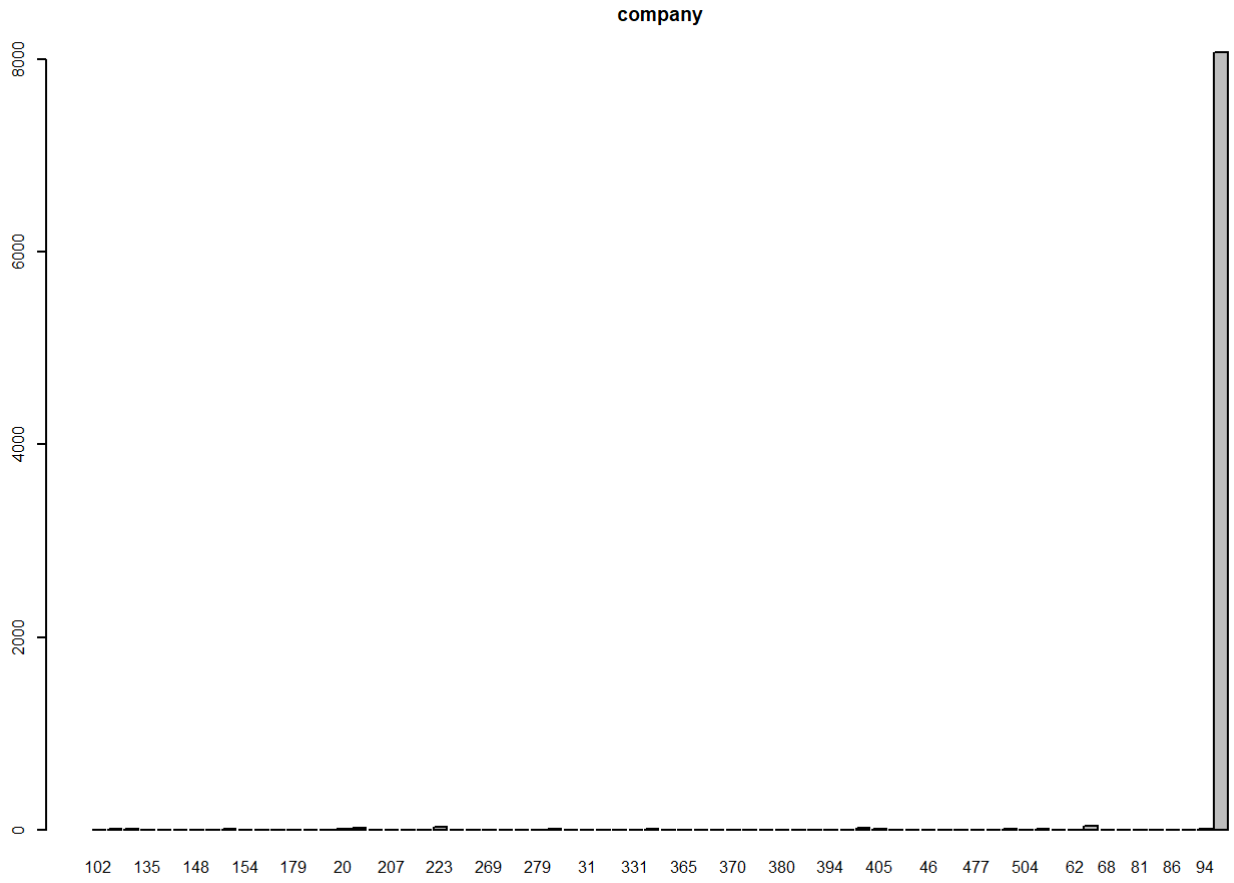

reserved_room_type
0
1000
2000
3000
4000
5000
6000
A
B
C
D
E
F
G
H
L
P

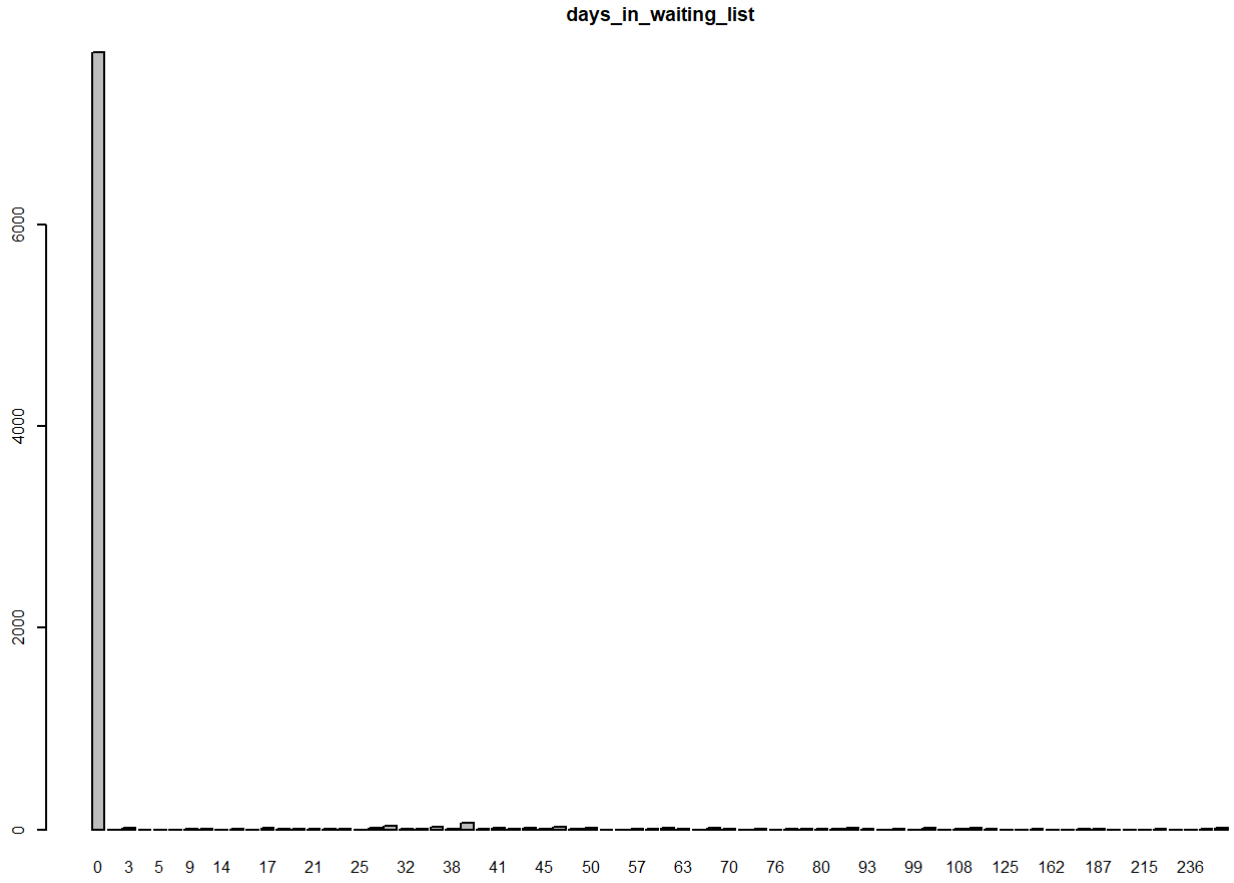

assigned_room_type
0
1000
2000
3000
4000
5000
6000
A
B
C
D
E
F
G
H
I
K
L
P

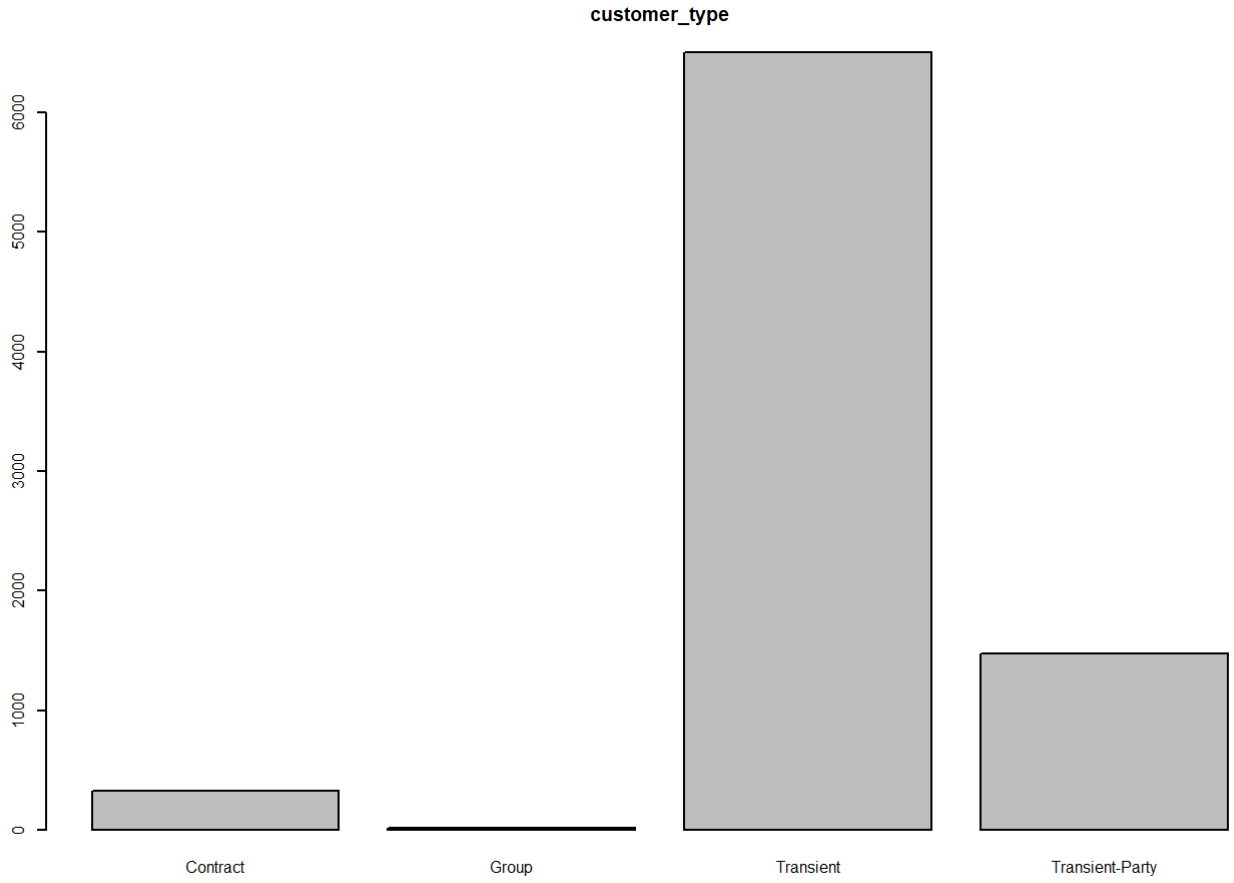

booking_changes
0
2000
4000
6000
0
1
2
3
4
6
7
8

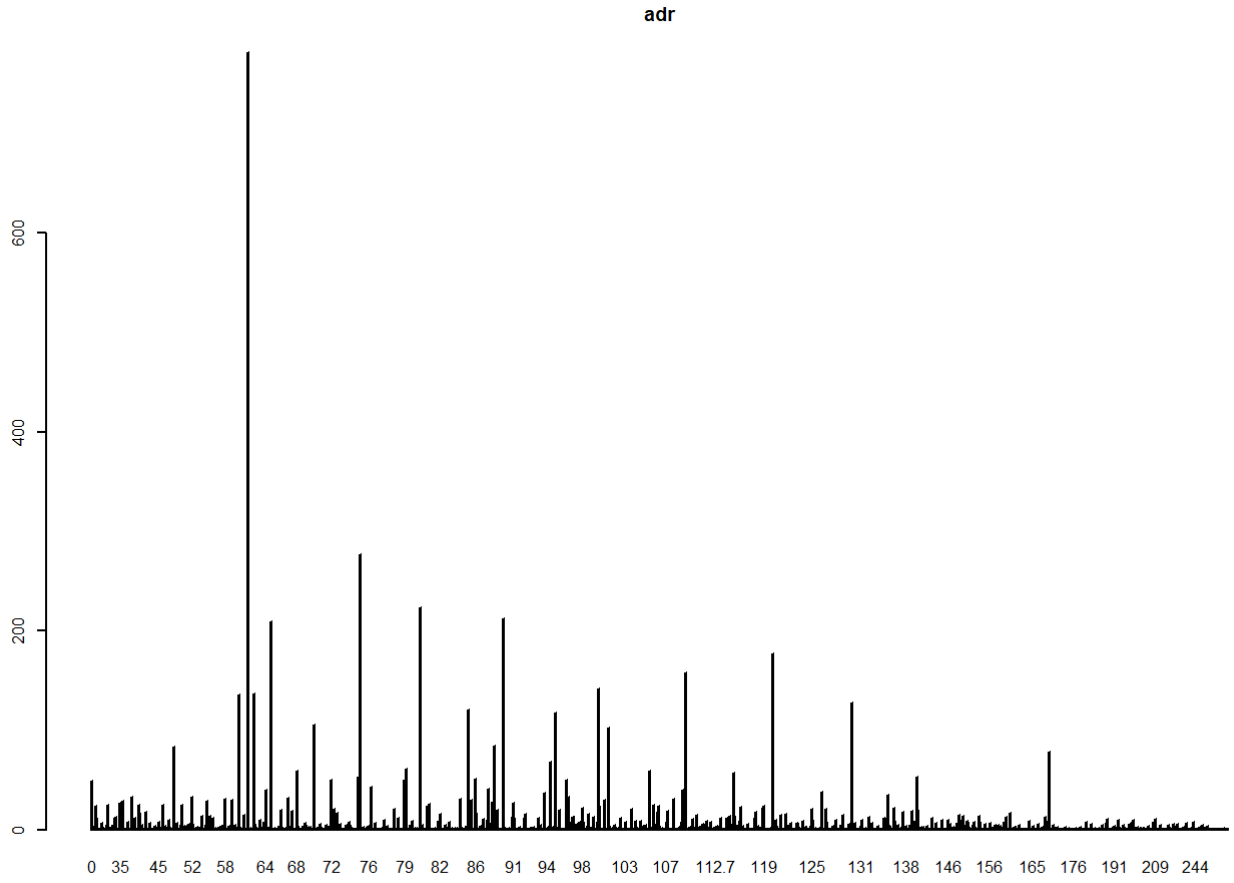

deposit_type
5000
4000
3000
2000
1000
0
No Deposit
Non Refund
Refundable

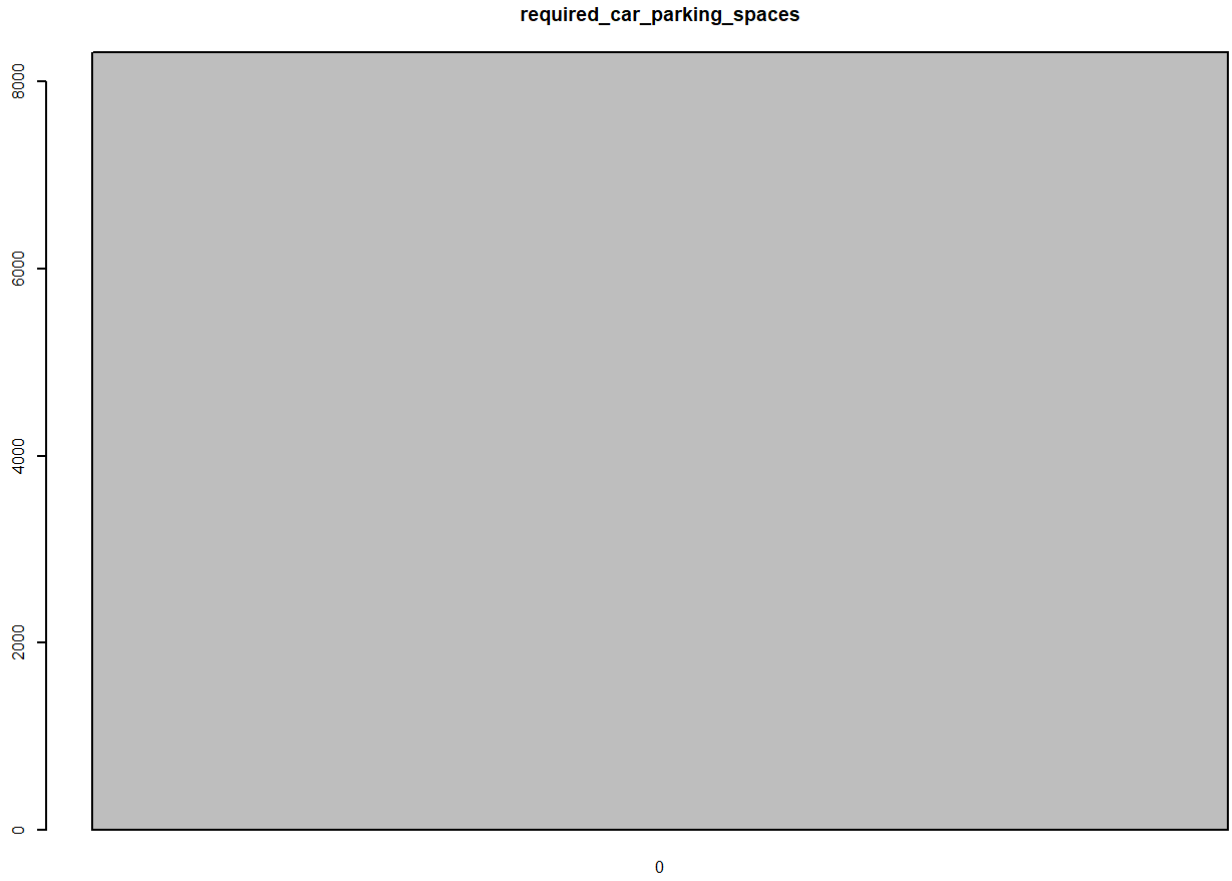

agent
2000
1500
1000
500
0
1 11 128 138 149 156 17 181 19 2 21 236 245 253 27 29 3 31 327 34 37 39 42 495 55 64 7 78 85 93 99

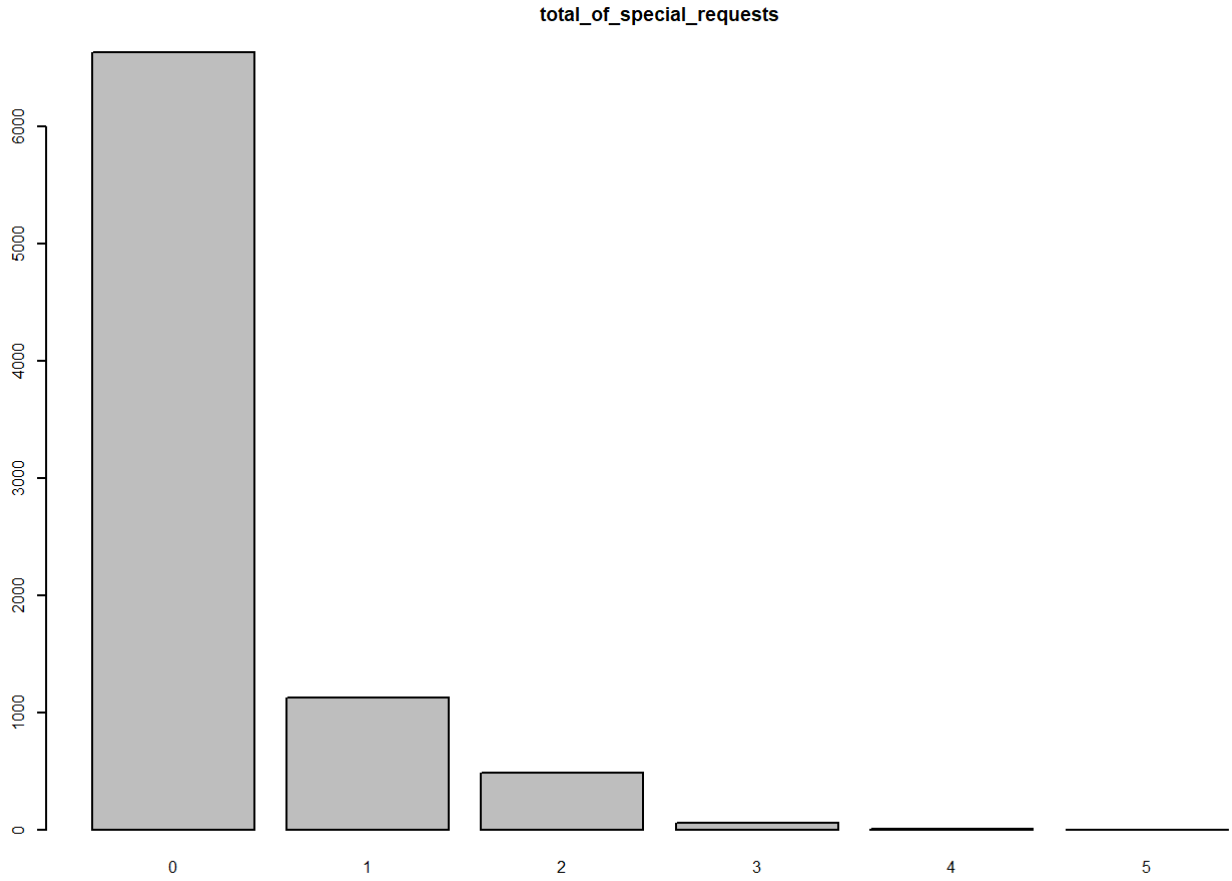

company
8000
6000
4000
2000
0
102 135 148 154 179 20 207 223 269 279 31 331 365 370 380 394 405 46 477 504 62 68 81 86 94


days_in_waiting_list
6000
4000
2000
0
0 3 5 9 14 17 21 25 32 38 41 45 50 57 63 70 76 80 93 99 108 125 162 187 215 236


customer_type
0
1000
2000
3000
4000
5000
6000
Contract
Group
Transient
Transient-Party


adr
0
200
400
600
0 35 45 52 58 64 68 72 76 79 82 86 91 94 98 103 107 112.7 119 125 131 138 146 156 165 176 191 209 244


required_car_parking_spaces
8000
6000
4000
2000
0
0

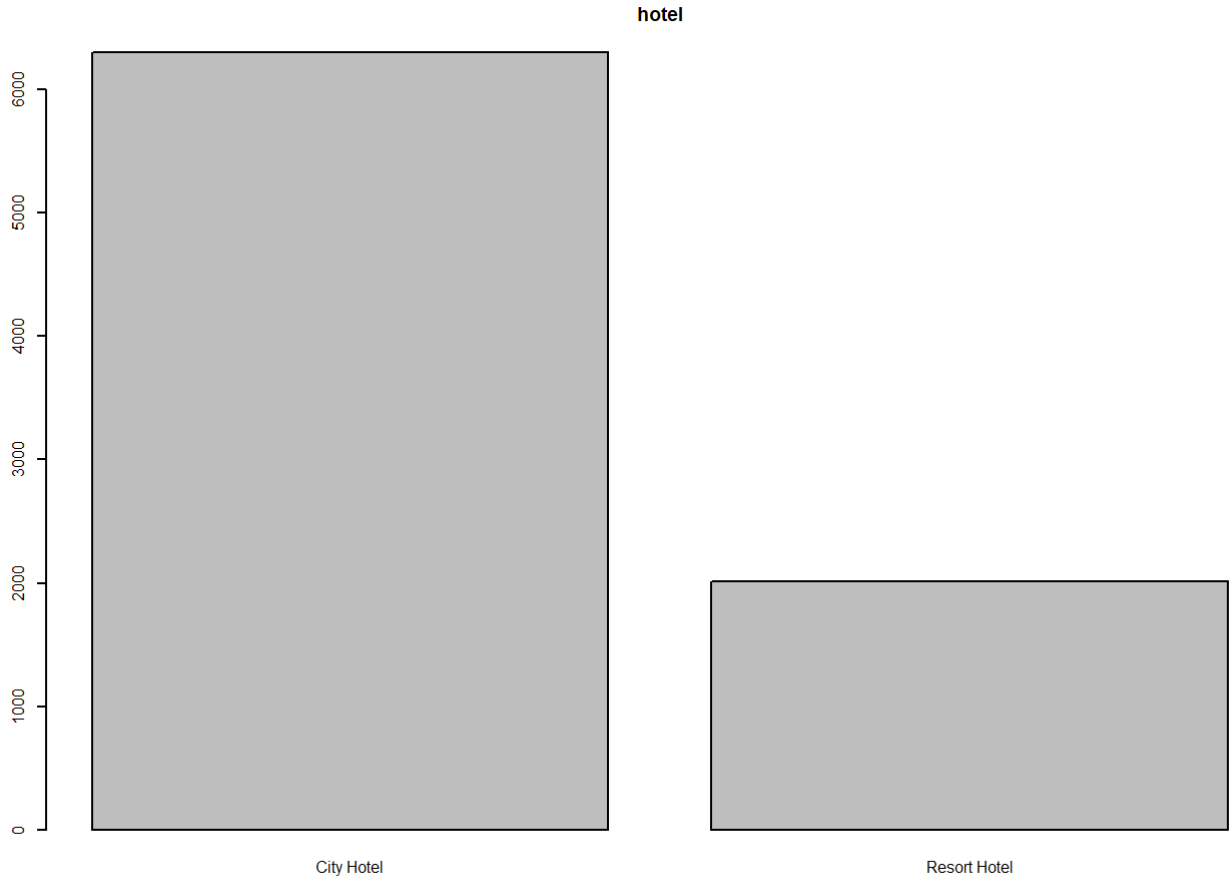

total_of_special_requests
6000
5000
4000
3000
2000
1000
0
0
1
2
3
4
5

# analisis de todas las variables con el subconjunto de canceladas ordenas

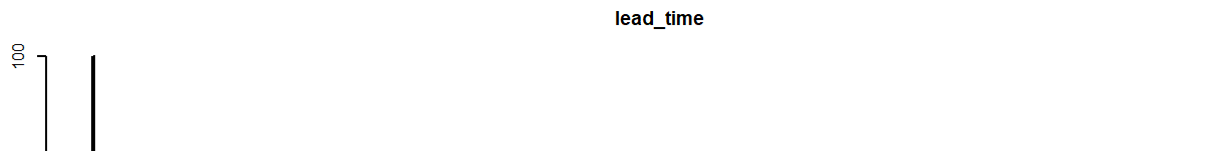
```
ncol(sub_cancelados)
```

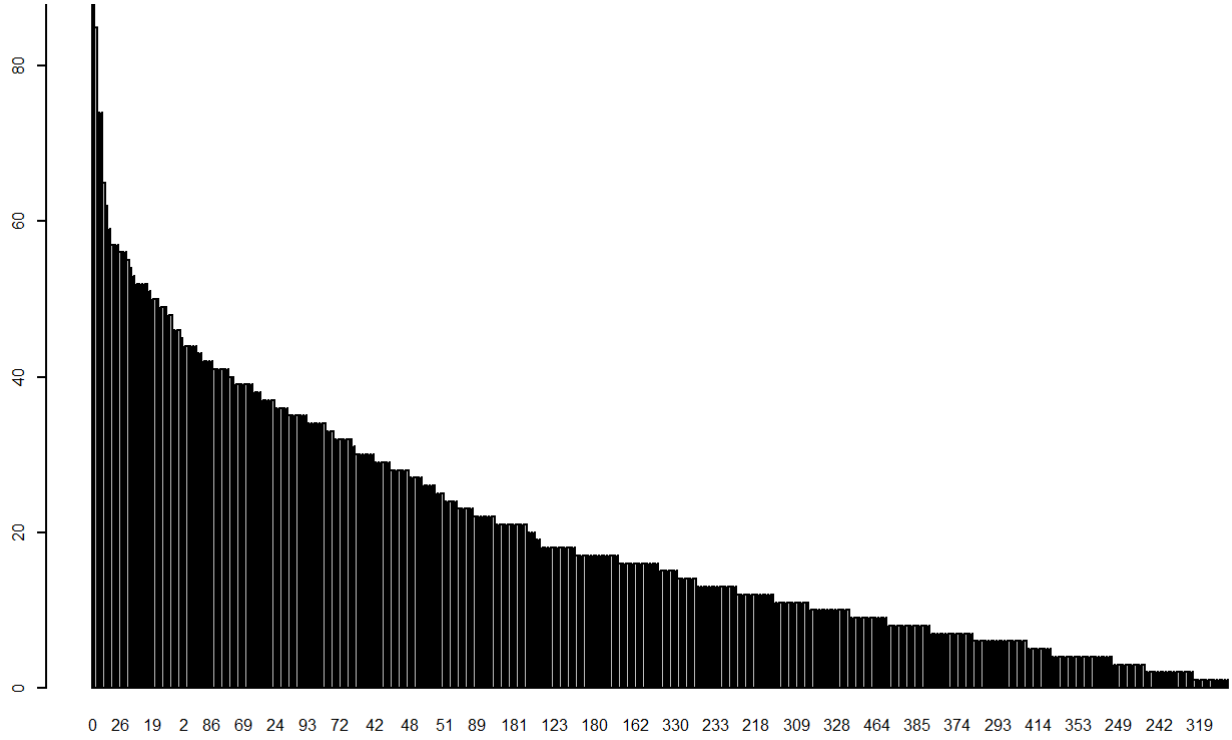
```
## [1] 30
```

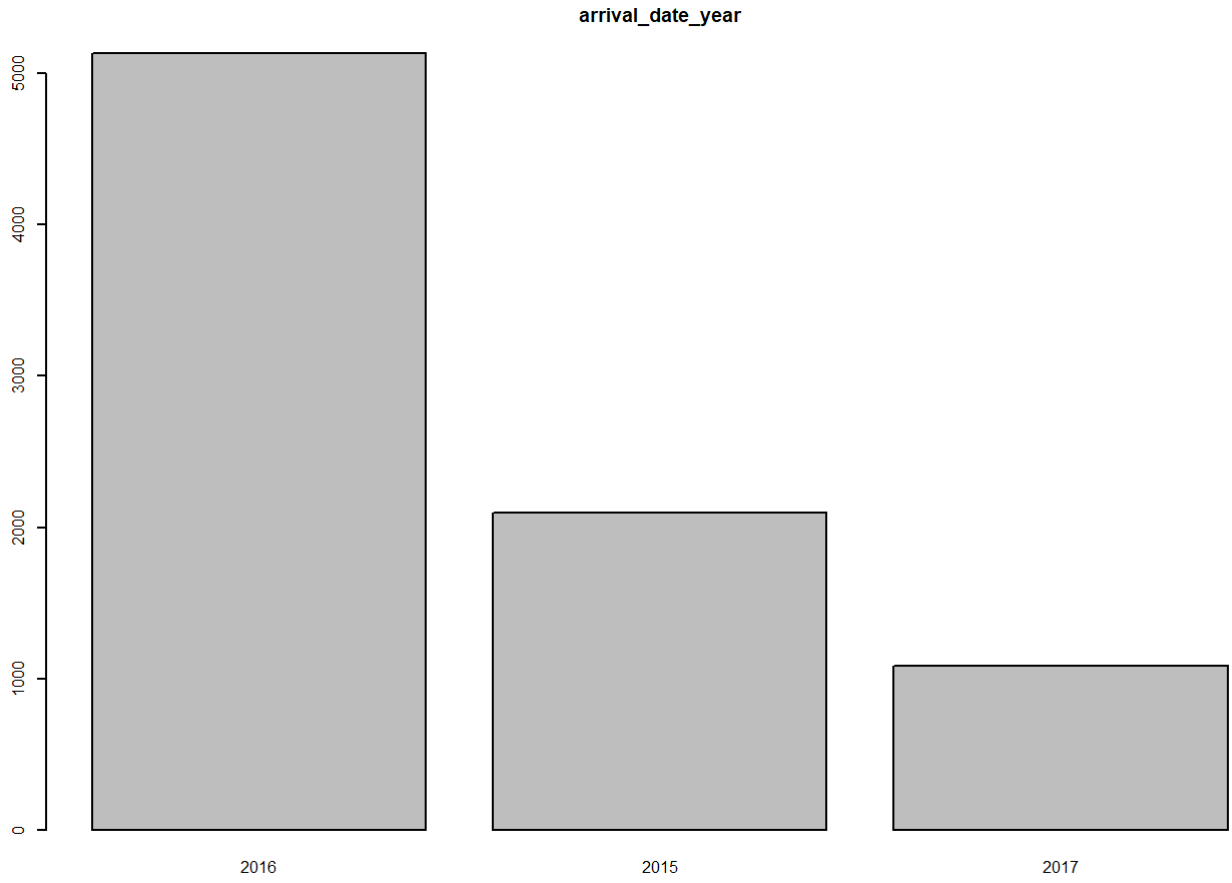
```
nombres <- names(sub_cancelados)
for (i in c(1:ncol(sub_cancelados))){
analisis_cancelaciones_ord(sub_cancelados[,i],nombres[i])
}
```

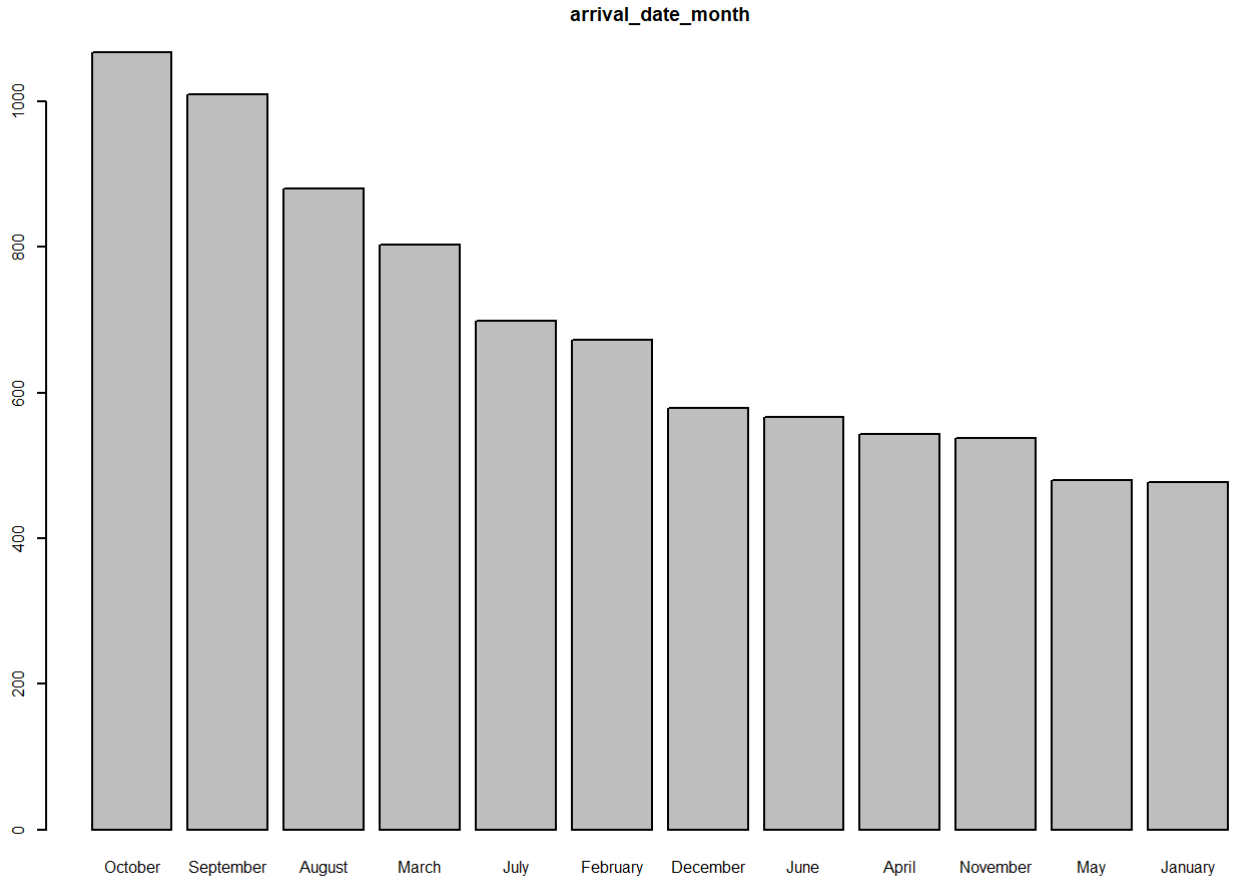

hotel
0
1000
2000
3000
4000
5000
6000
City Hotel
Resort Hotel

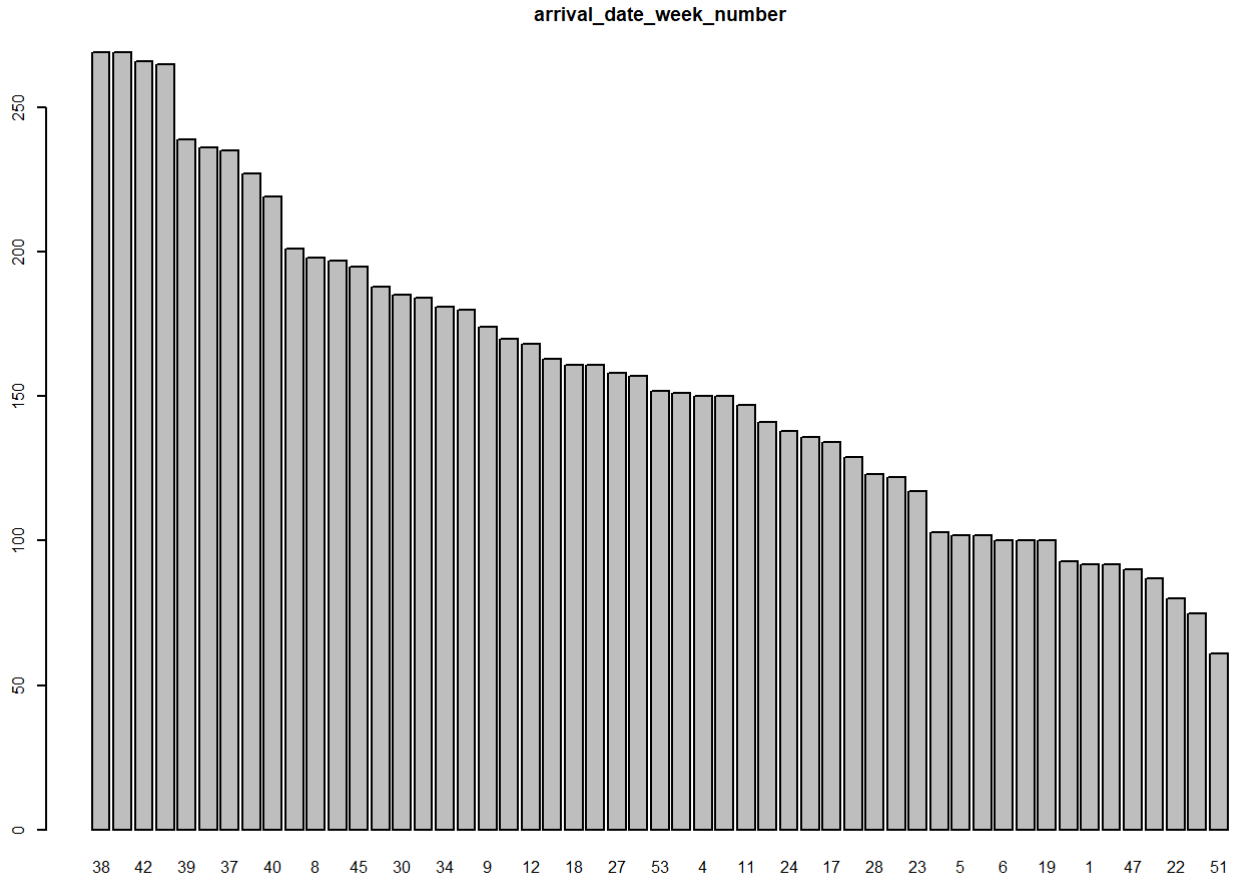

is_canceled
0
2000
4000
6000
8000
cancelado

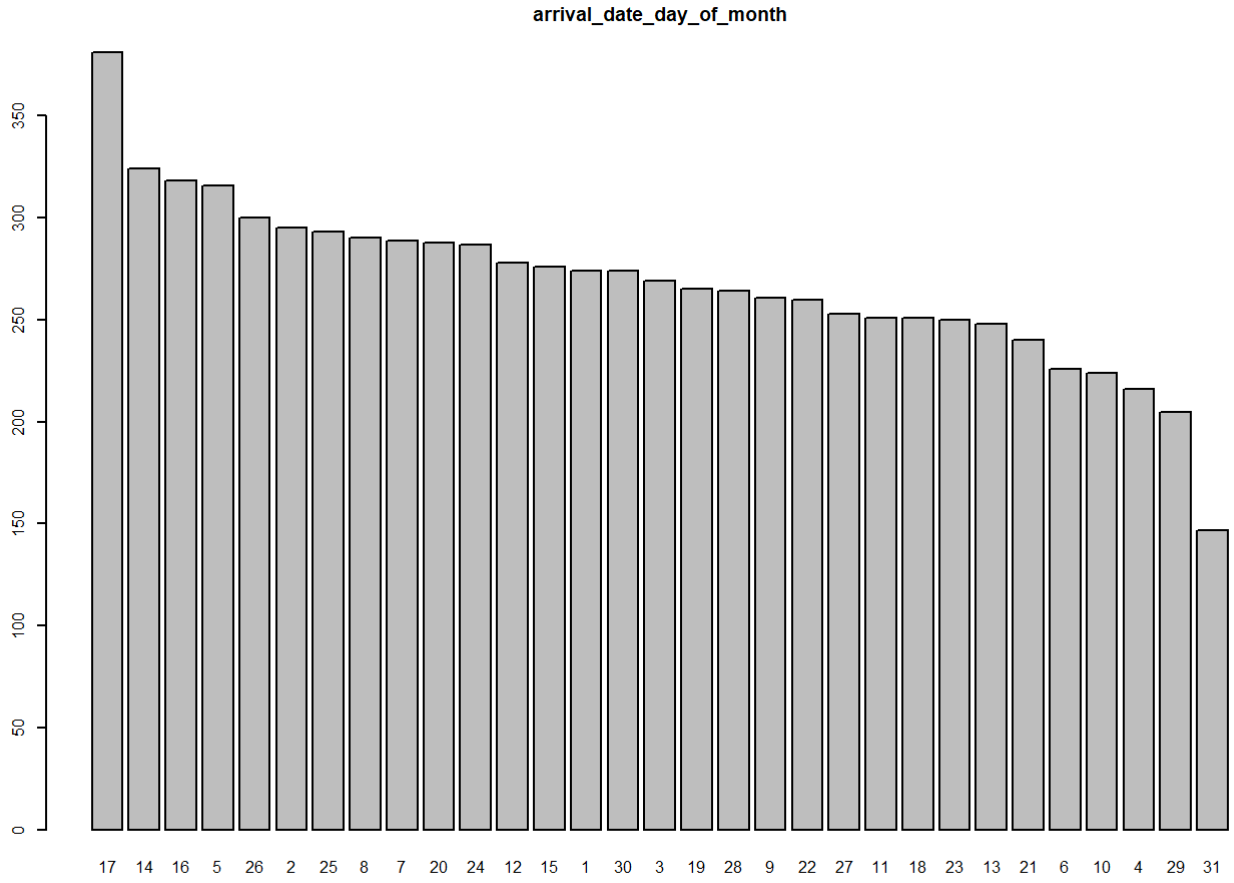

lead_time
100

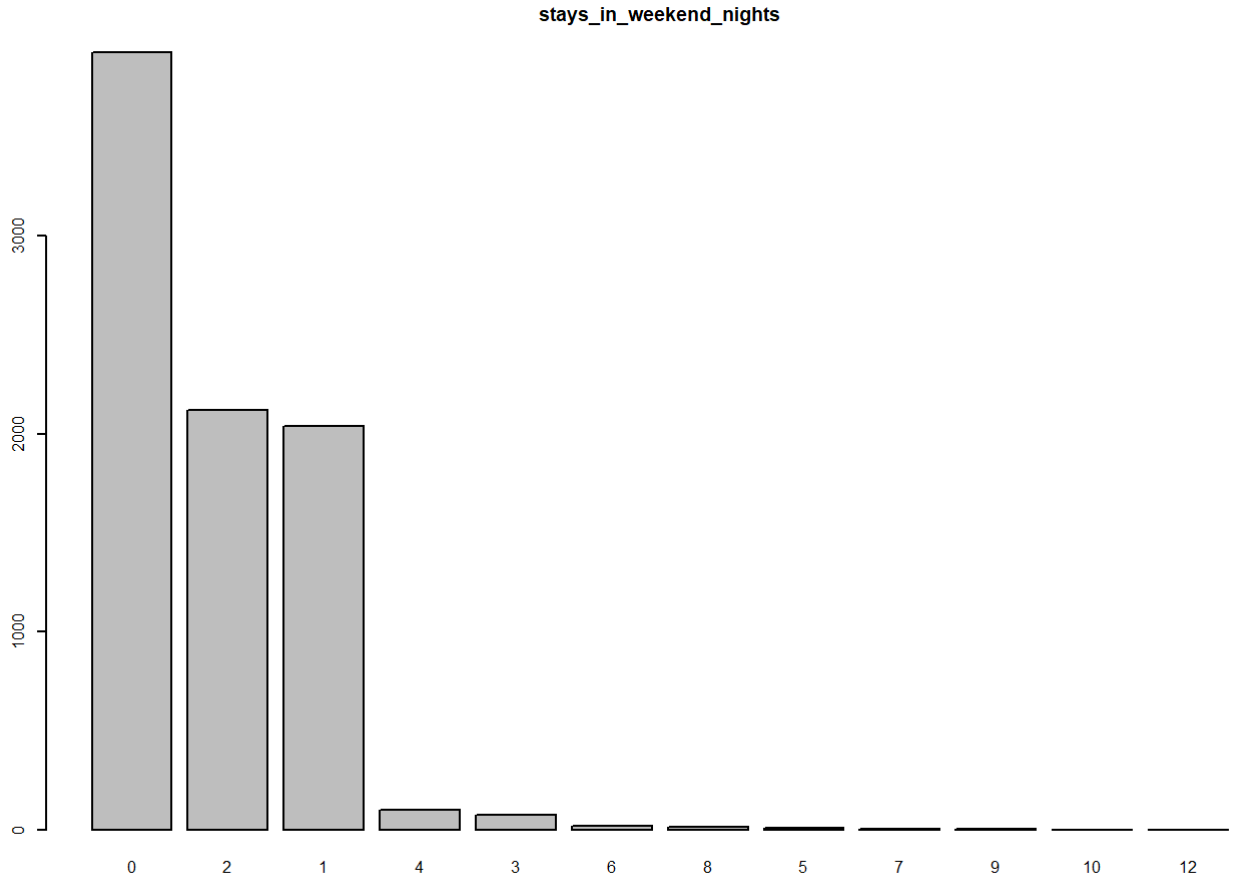

0
20
40
60
80
0 26 19 2 86 69 24 93 72 42 48 51 89 181 123 180 162 330 233 218 309 328 464 385 374 293 414 353 249 242 319

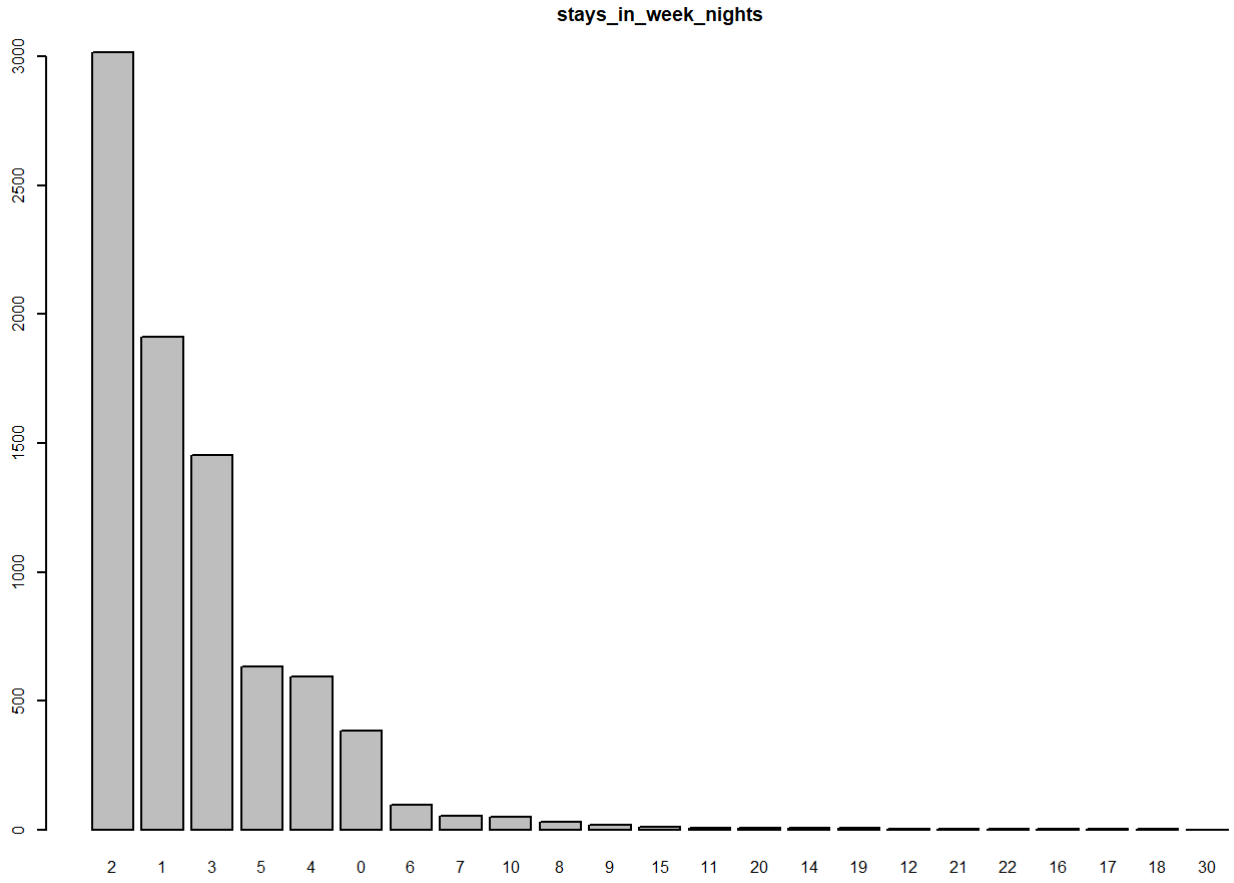

arrival_date_year
0
1000
2000
3000
4000
5000
2016
2015
2017

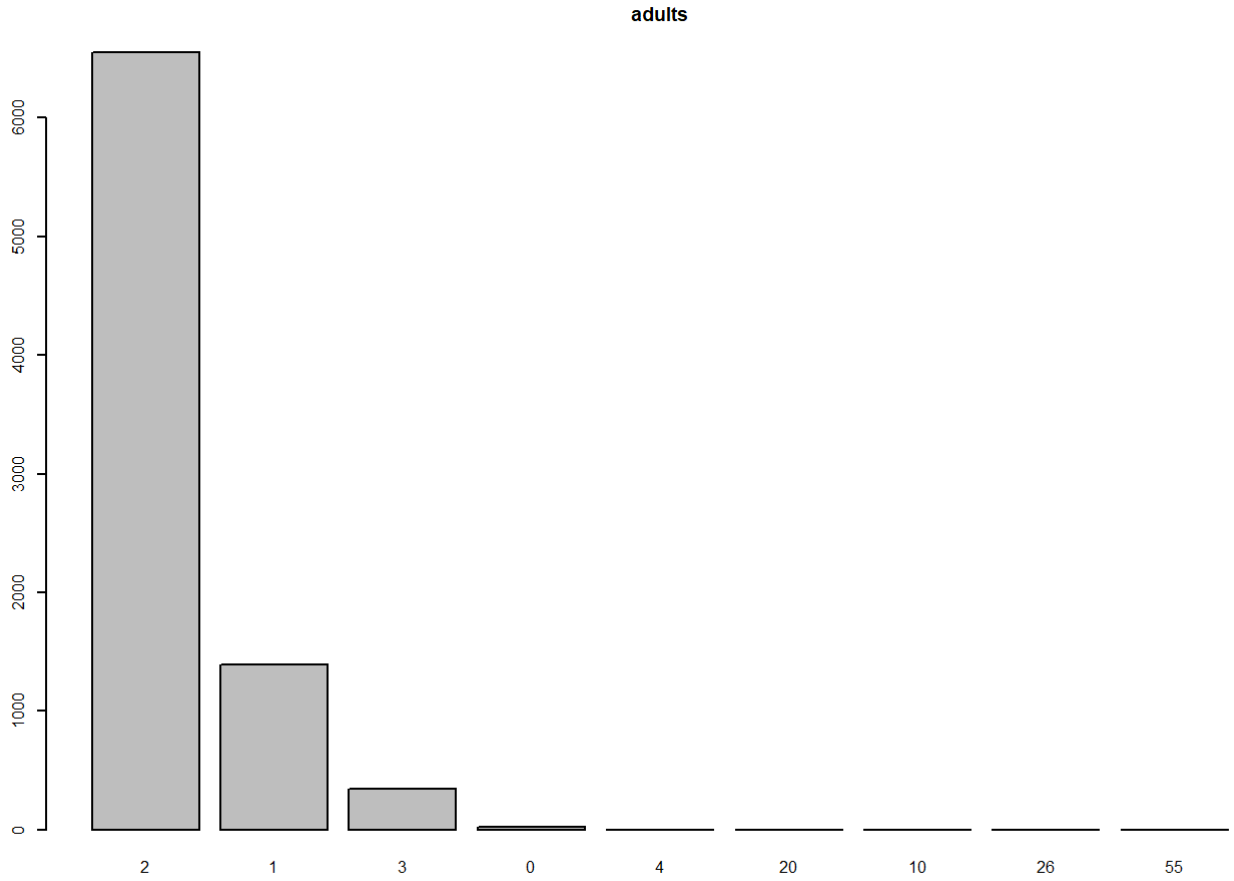

arrival_date_month
0
200
400
600
800
1000
October
September
August
March
July
February
December
June
April
November
May
January

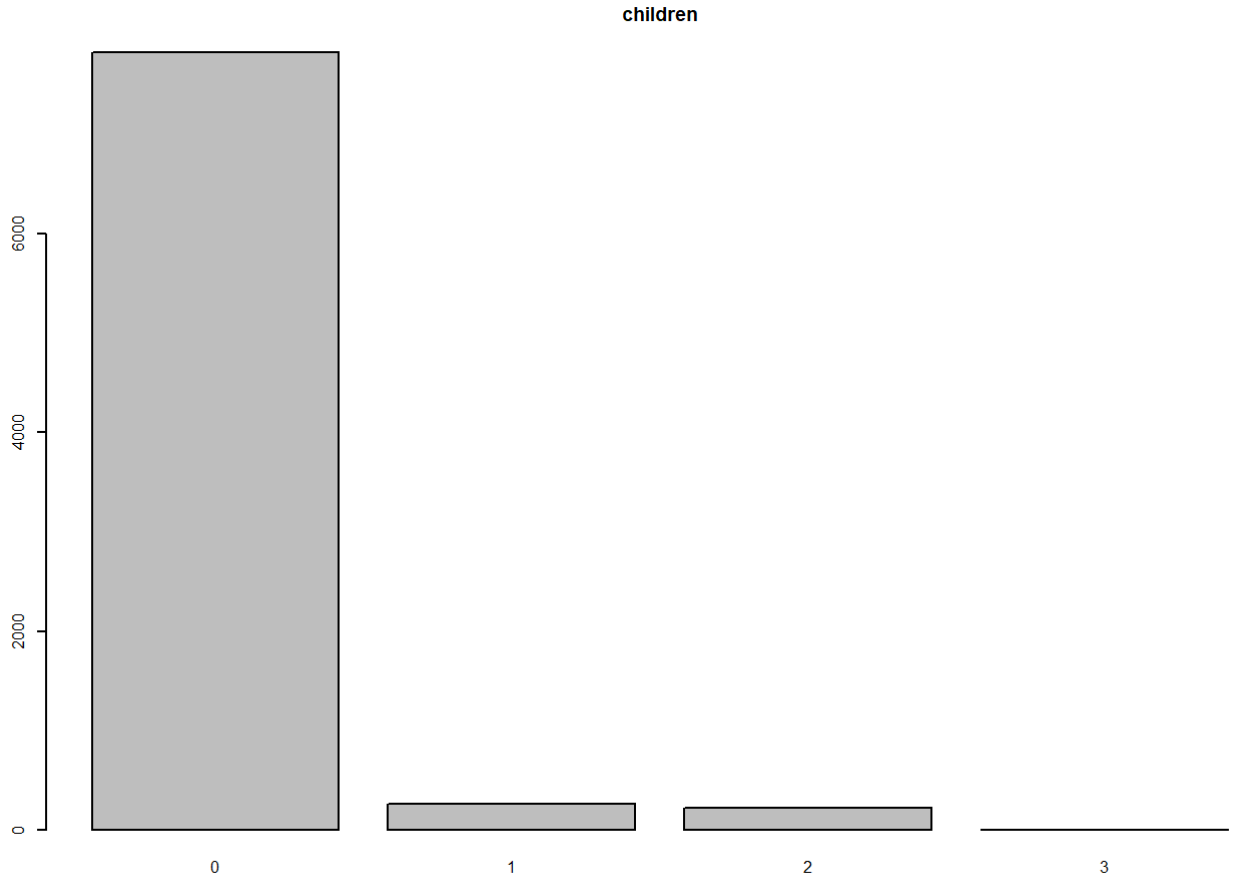

arrival_date_week_number
0
50
100
150
200
250
38
42
39
37
40
8
45
30
34
9
12
18
27
53
4
11
24
17
28
23
5
6
19
1
47
22
51

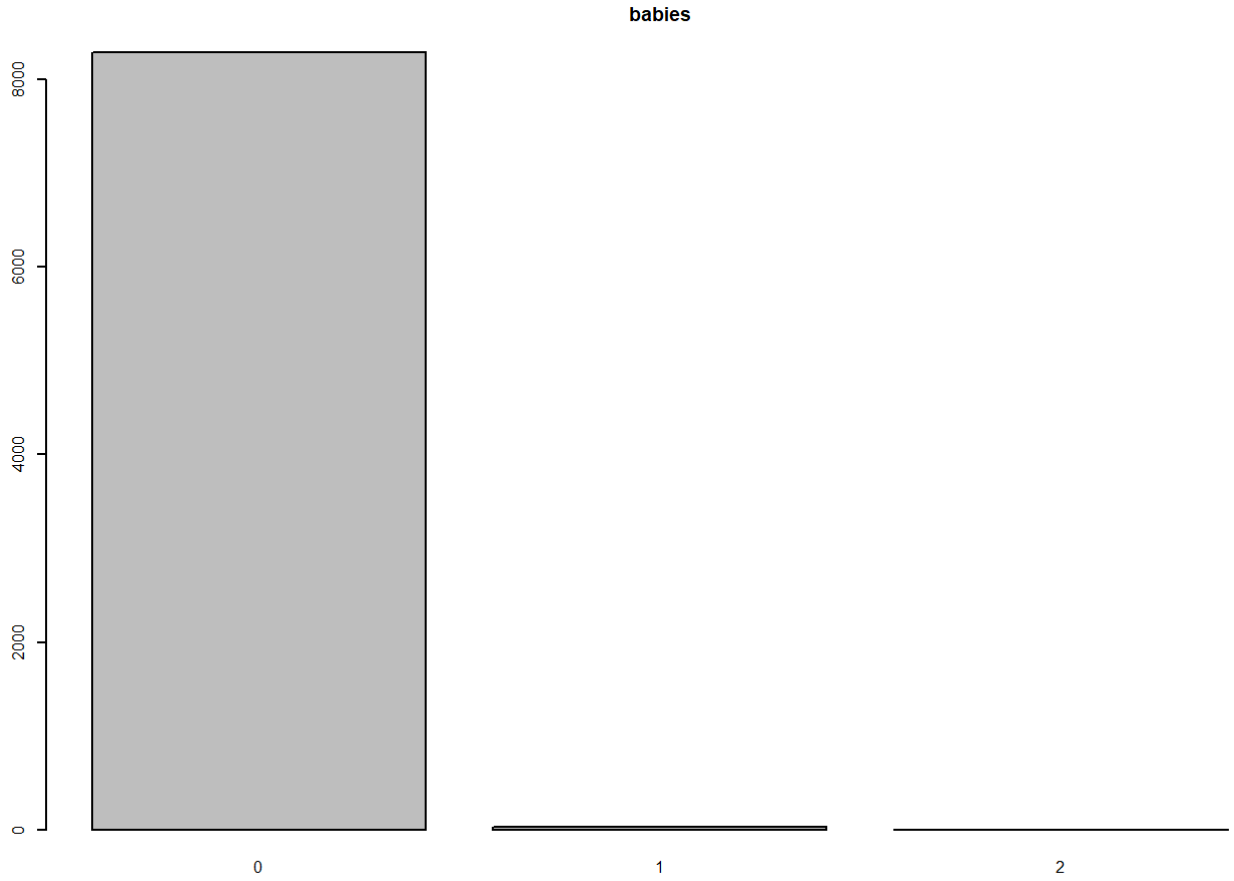

arrival_date_day_of_month
0
50
100
150
200
250
300
350
17 14 16 5 26 2 25 8 7 20 24 12 15 1 30 3 19 28 9 22 27 11 18 23 13 21 6 10 4 29 31

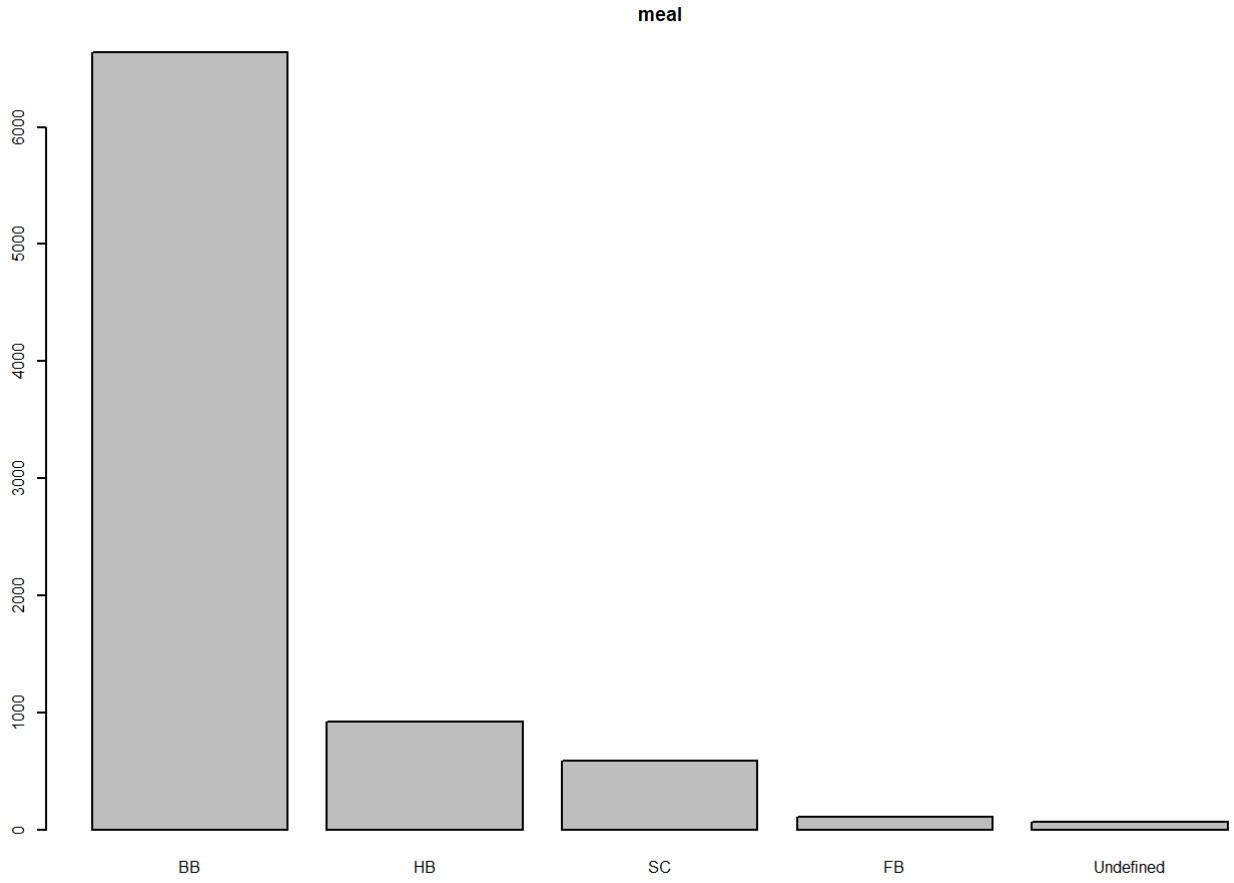

stays_in_weekend_nights
0
1000
2000
3000
0 2 1 4 3 6 8 5 7 9 10 12

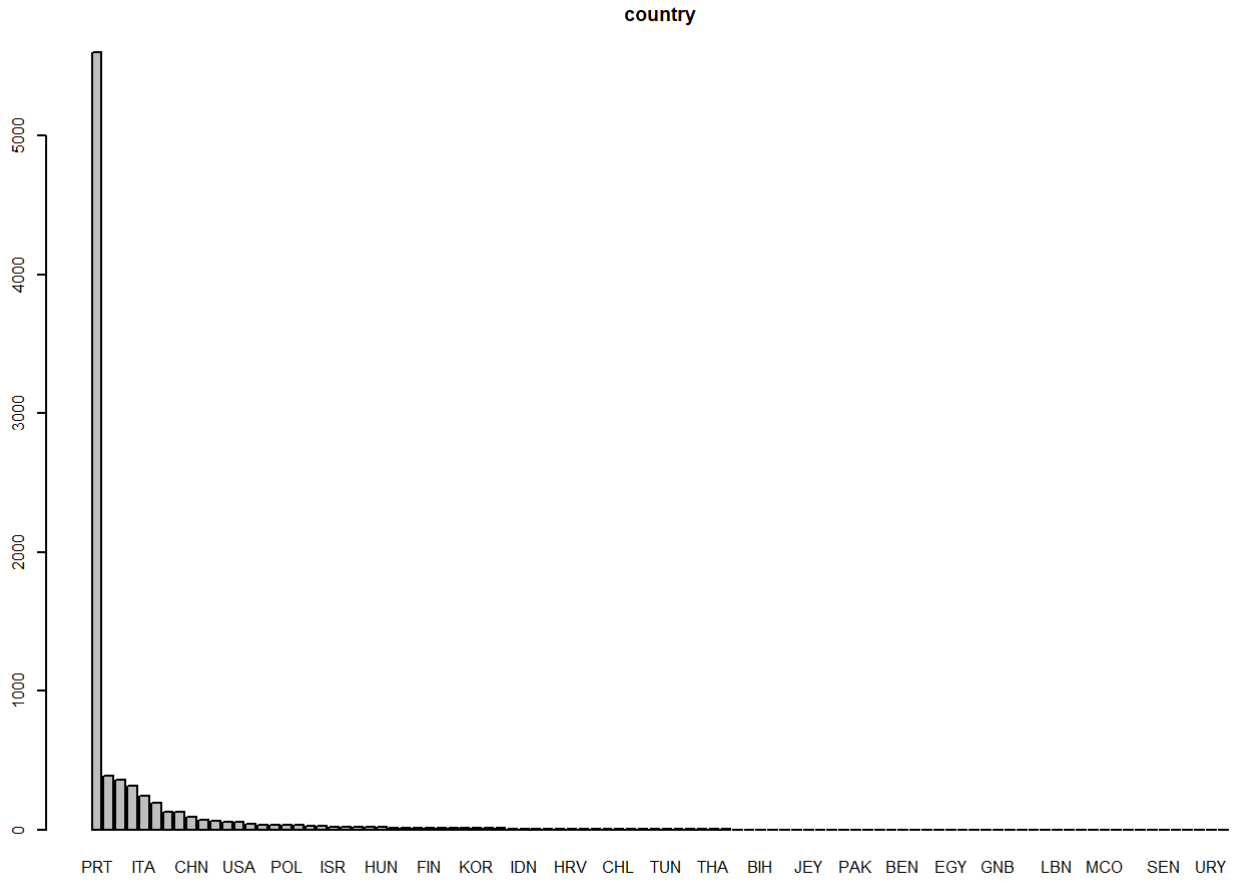

stays_in_week_nights
0
500
1000
1500
2000
2500
3000
2 1 3 5 4 0 6 7 10 8 9 15 11 20 14 19 12 21 22 16 17 18 30

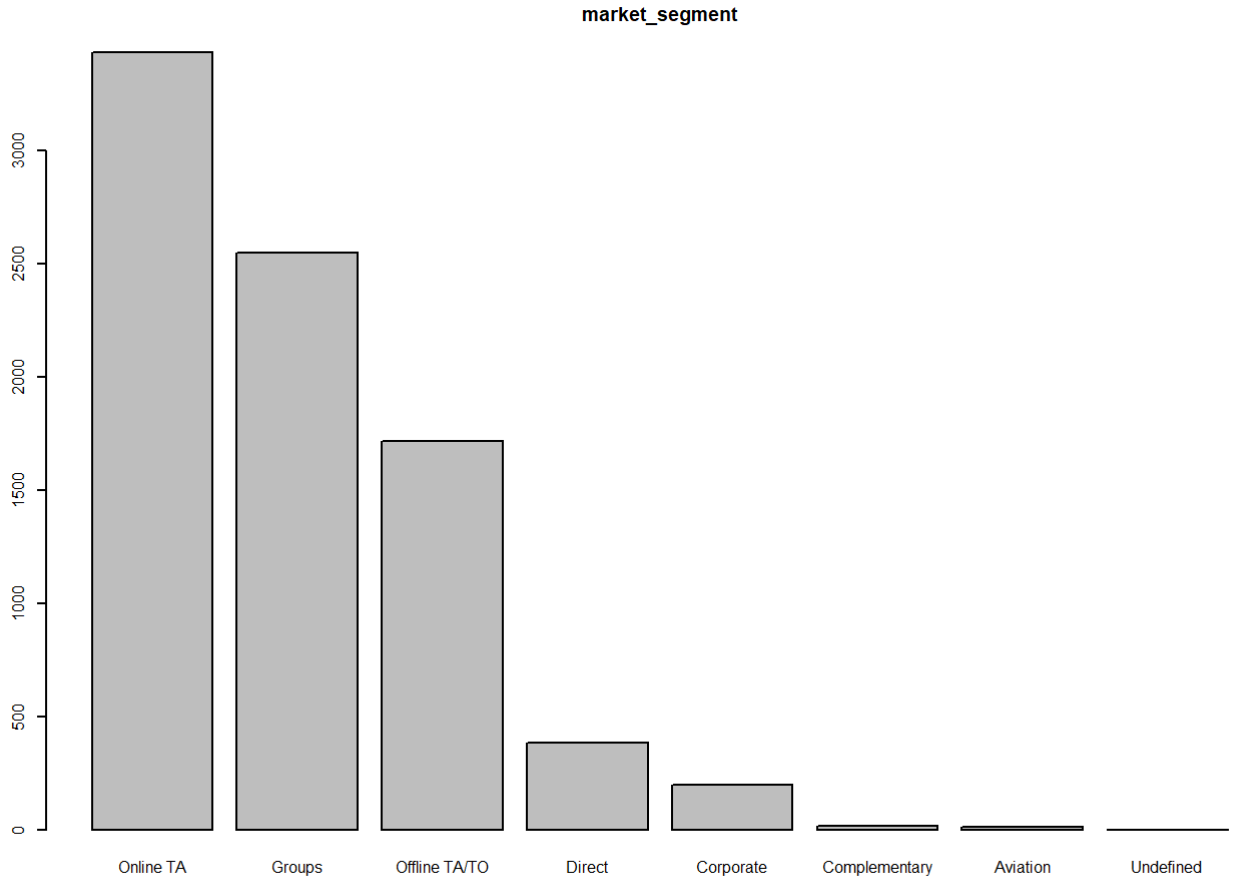

adults
0
1000
2000
3000
4000
5000
6000
2 1 3 0 4 20 10 26 55

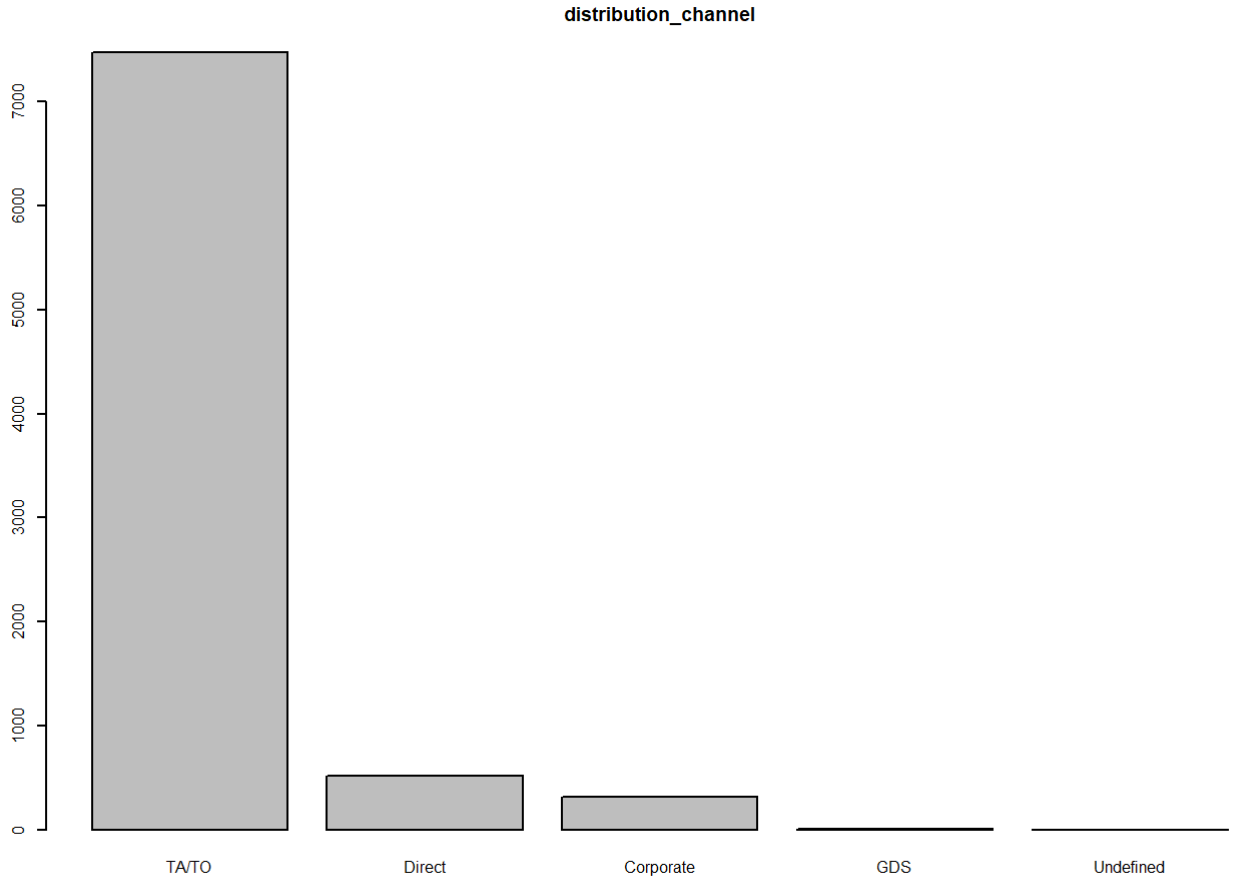

children
6000
4000
2000
0
0
1
2
3

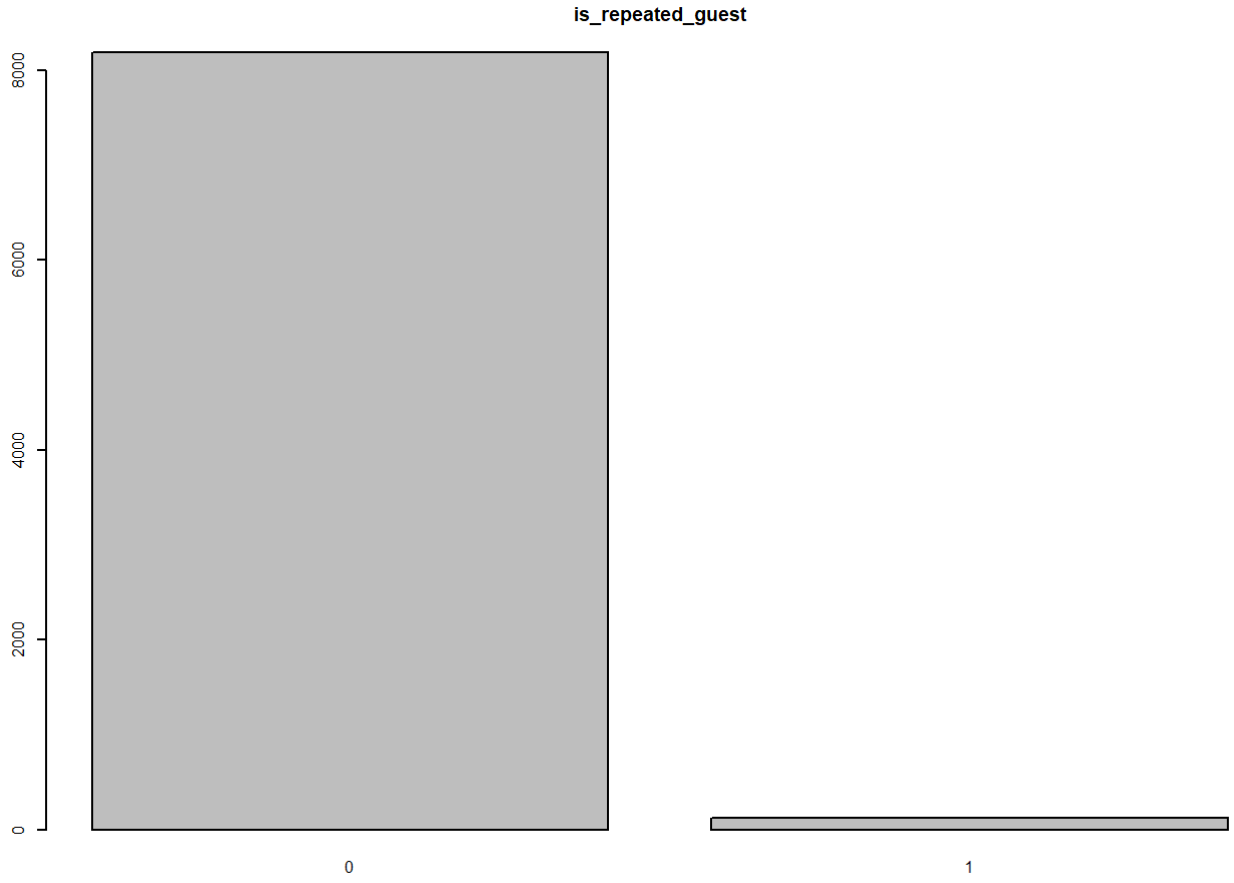

babies
8000
6000
4000
2000
0
0
1
2

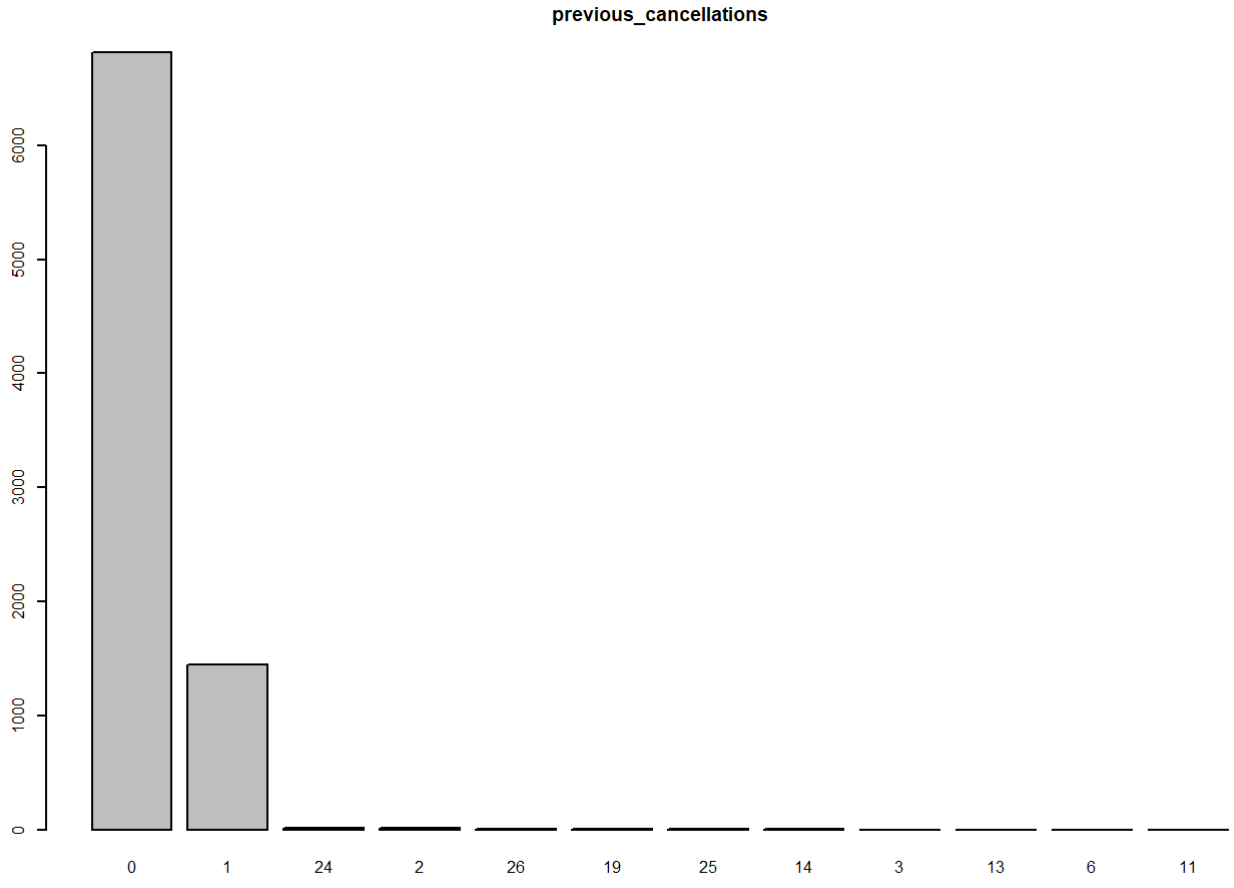

meal
0
1000
2000
3000
4000
5000
6000
BB
HB
SC
FB
Undefined

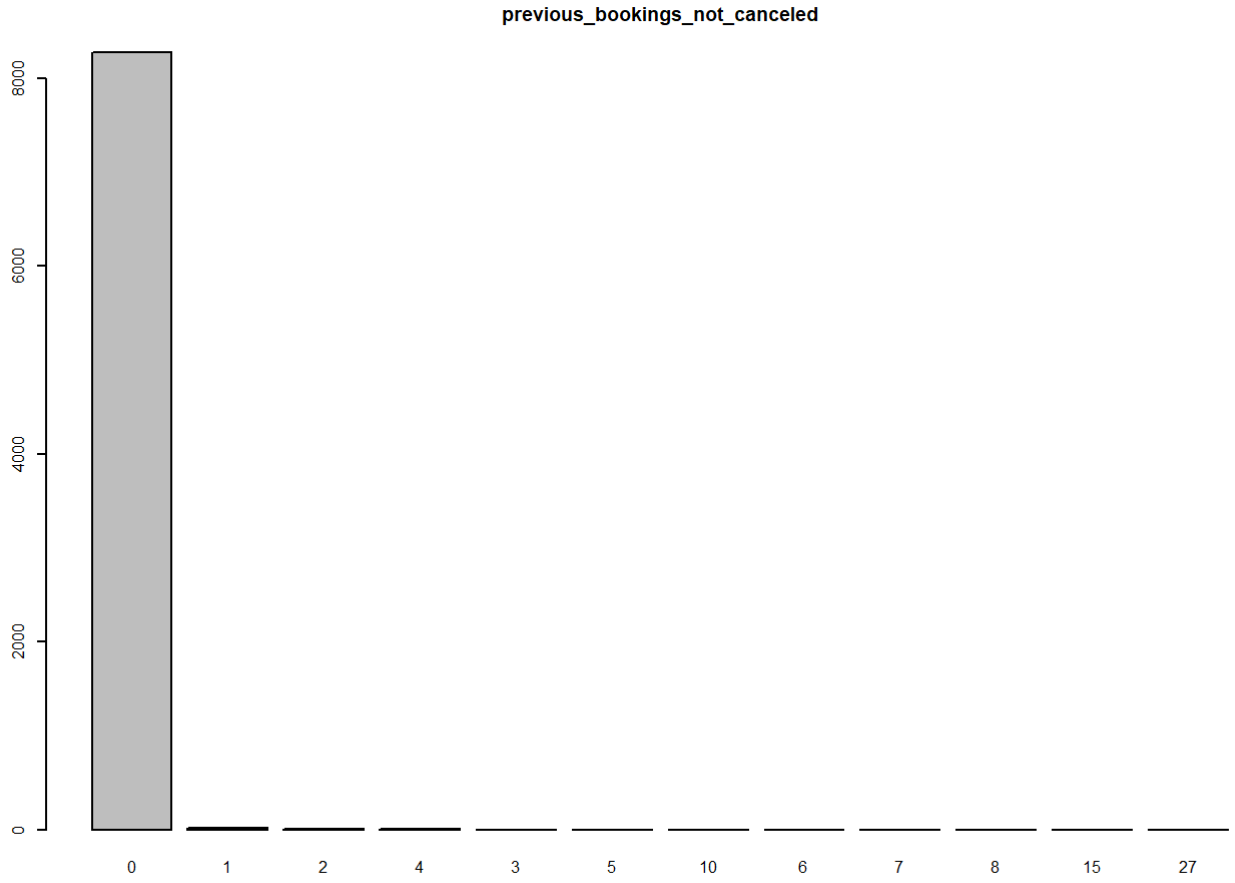

country
0
1000
2000
3000
4000
5000
PRT
ITA
CHN
USA
POL
ISR
HUN
FIN
KOR
IDN
HRV
CHL
TUN
THA
BIH
JEY
PAK
BEN
EGY
GNB
LBN
MCO
SEN
URY

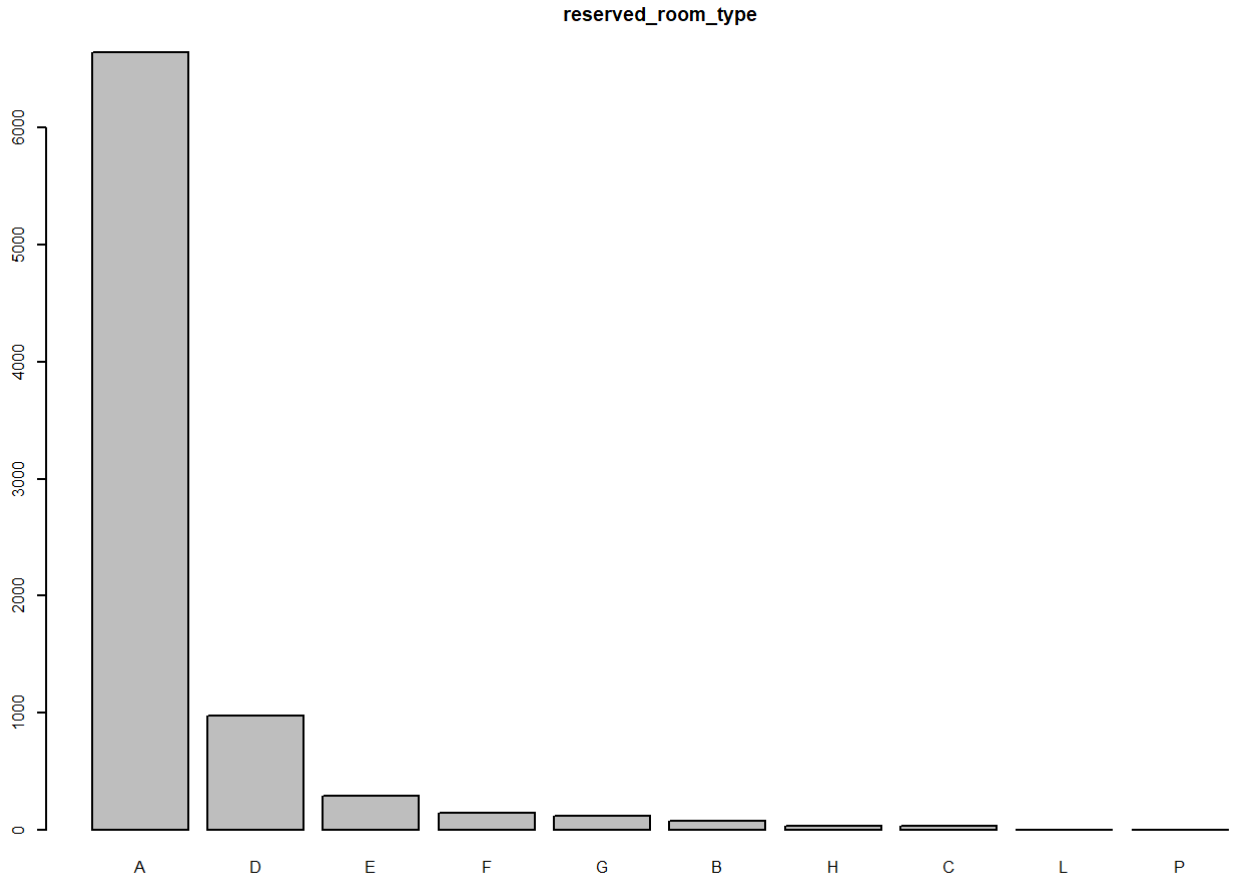

market_segment
0
500
1000
1500
2000
2500
3000
Online TA
Groups
Offline TA/TO
Direct
Corporate
Complementary
Aviation
Undefined

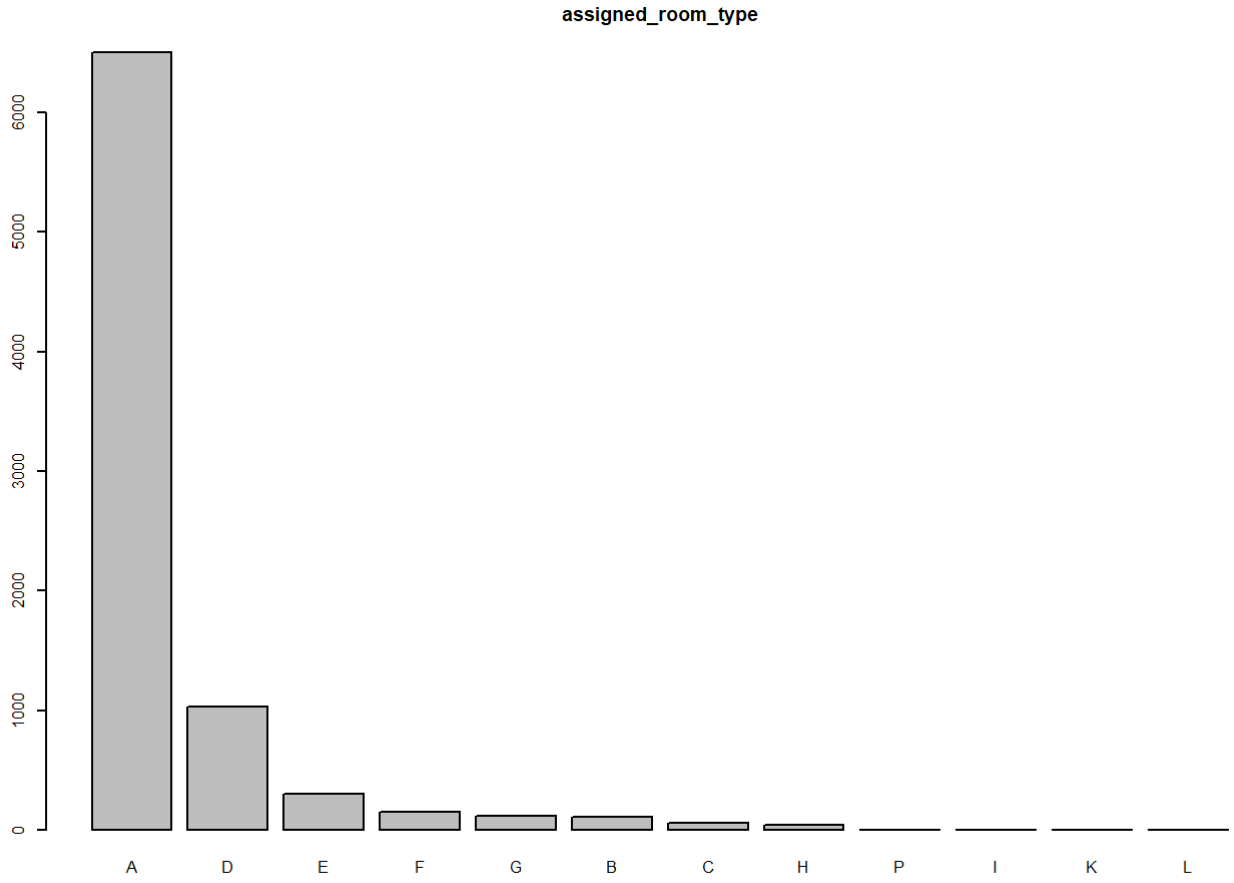

distribution_channel
0
1000
2000
3000
4000
5000
6000
7000
TA/TO
Direct
Corporate
GDS
Undefined

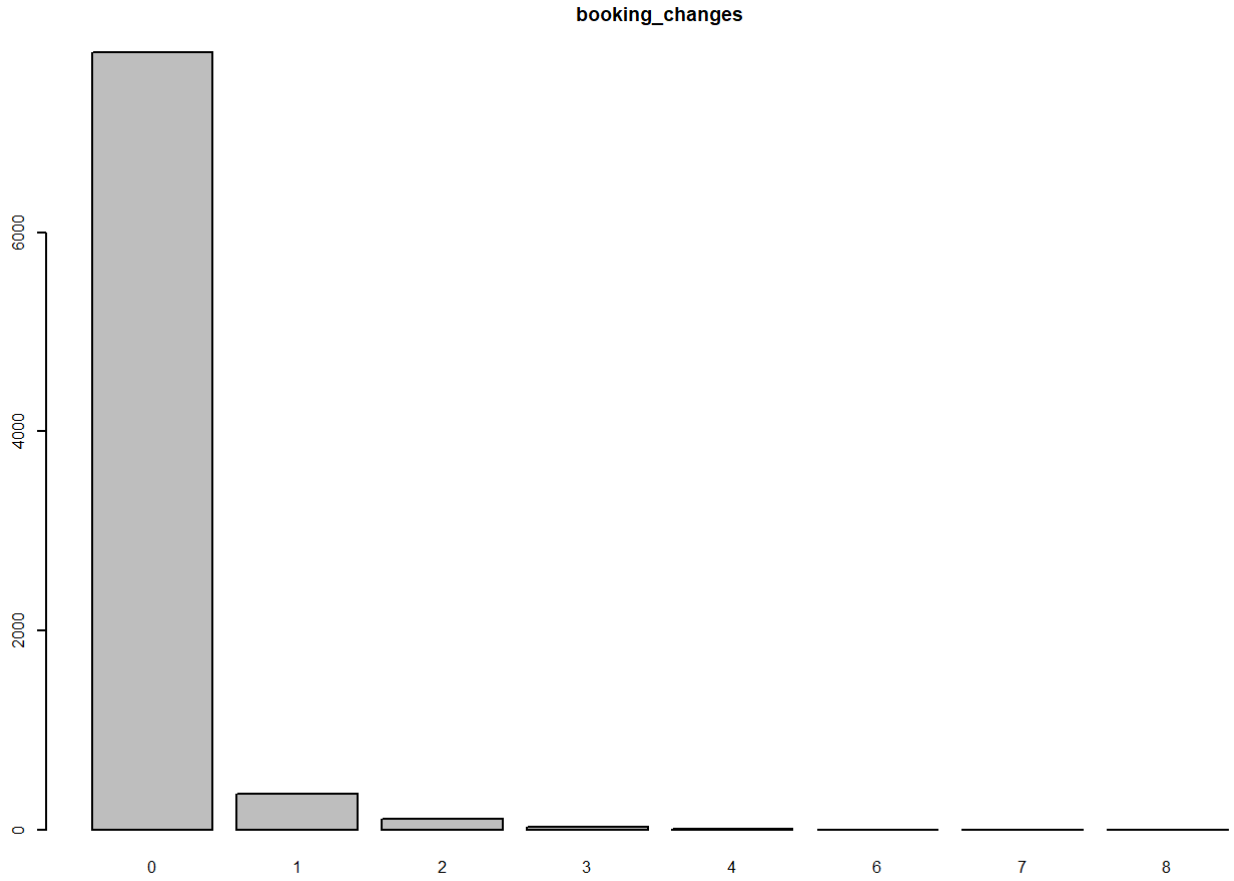

is_repeated_guest
8000
6000
4000
2000
0
0
1

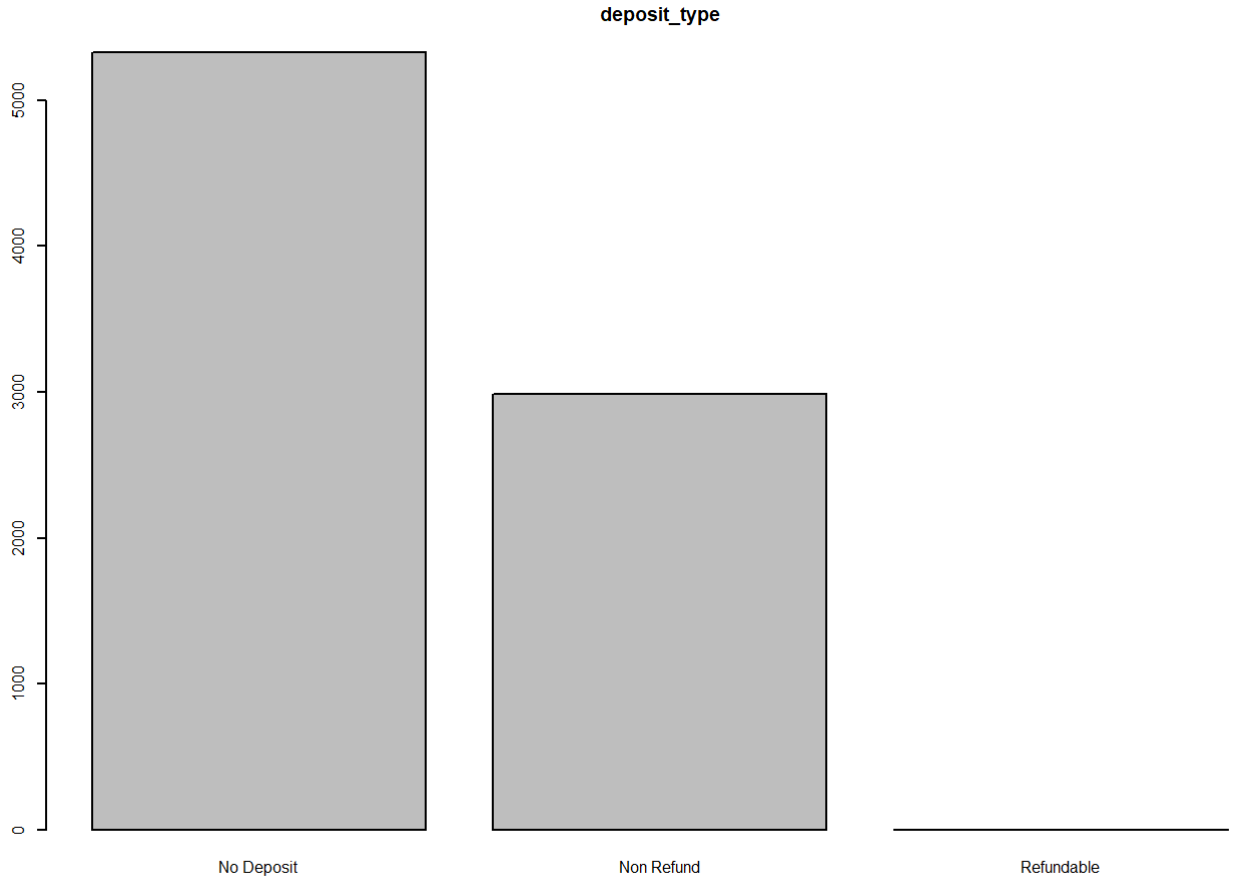

previous_cancellations
6000
5000
4000
3000
2000
1000
0
0
1
24
2
26
19
25
14
3
13
6
11

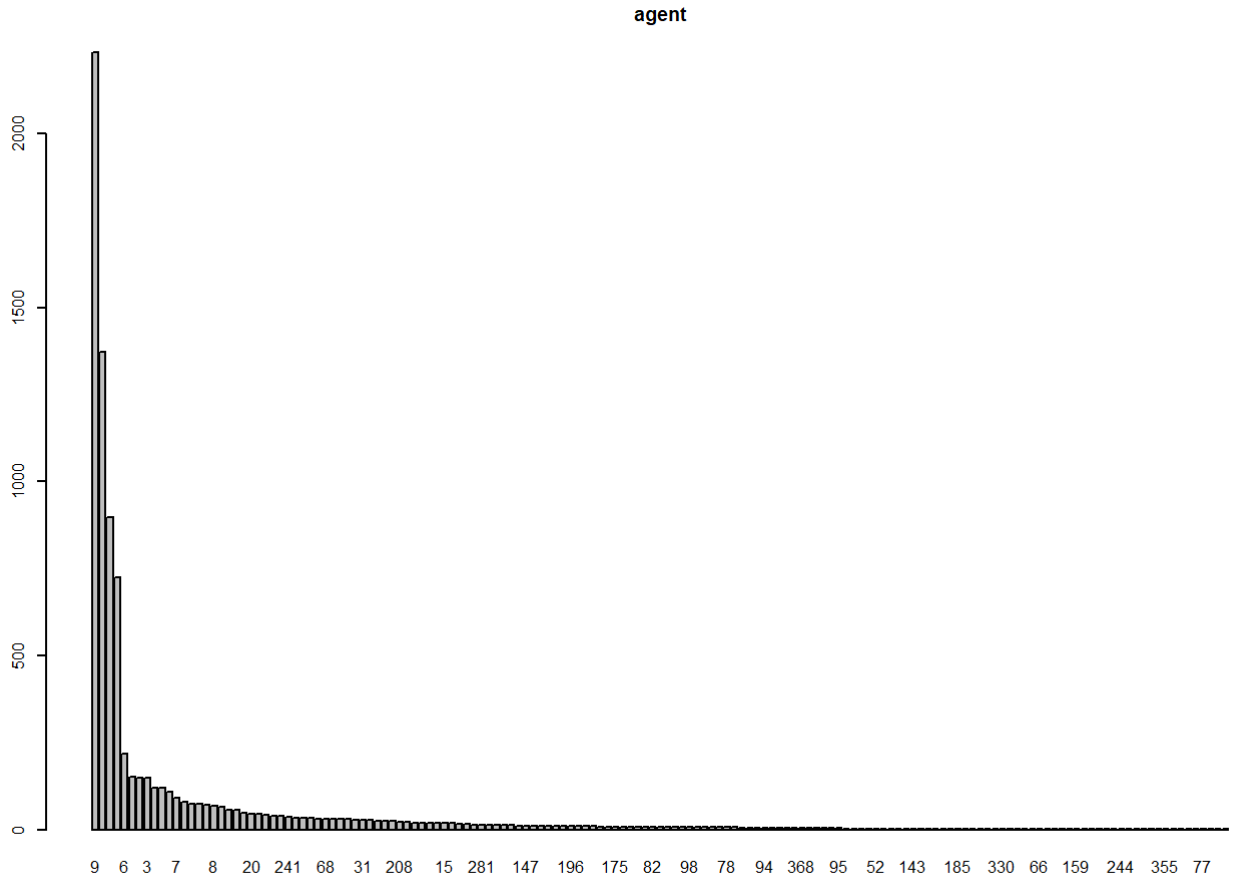

previous_bookings_not_canceled
0
2000
4000
6000
8000
0
1
2
4
3
5
10
6
7
8
15
27

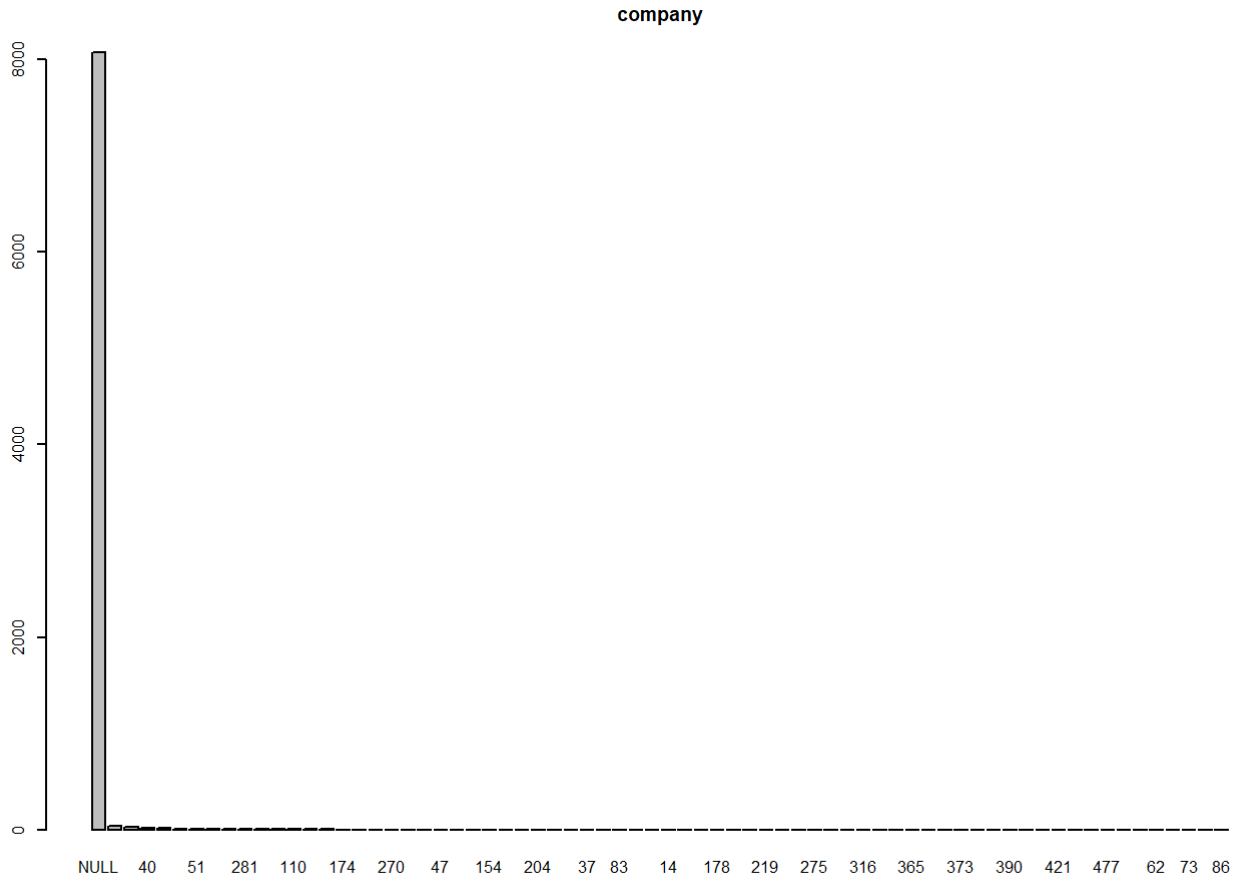

reserved_room_type
0
1000
2000
3000
4000
5000
6000
A
D
E
F
G
B
H
C
L
P

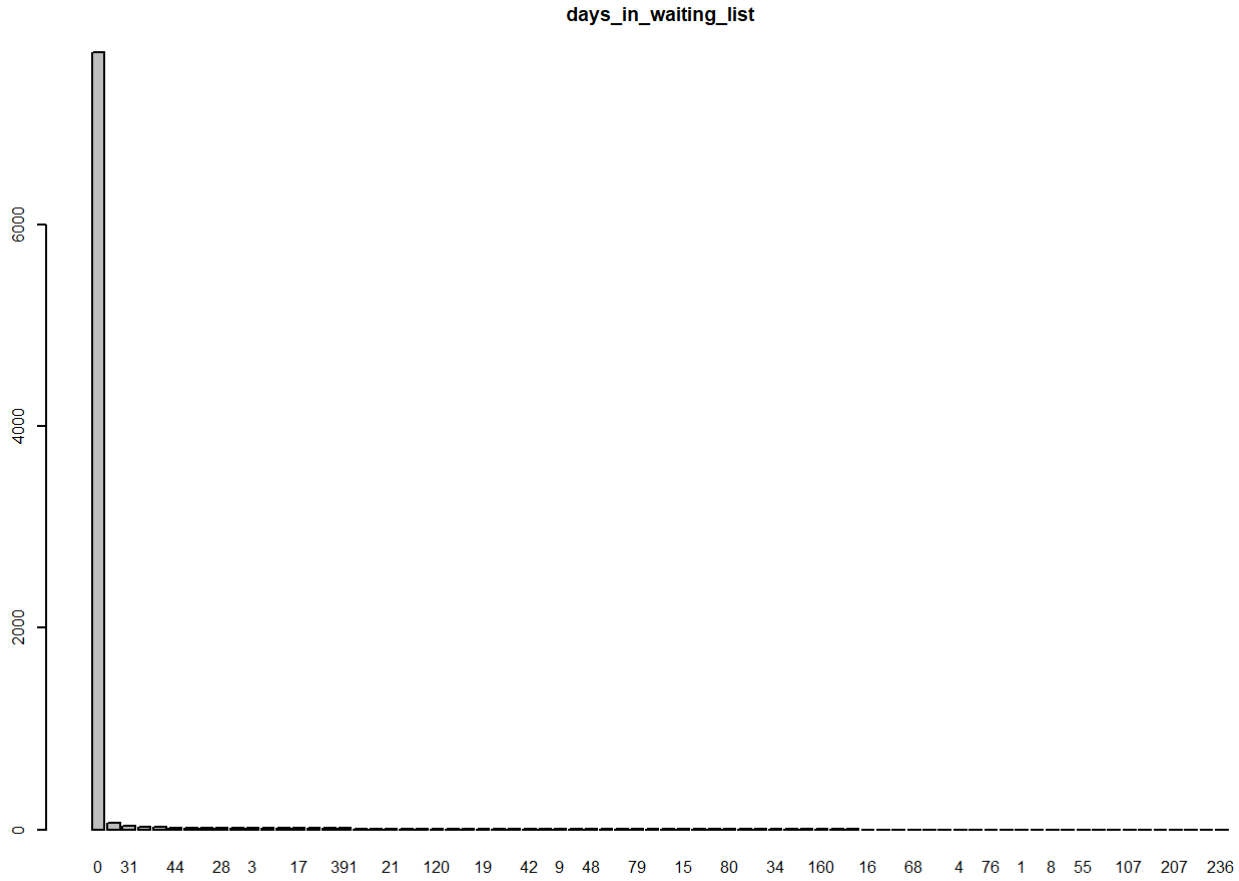

assigned_room_type
0
1000
2000
3000
4000
5000
6000
A
D
E
F
G
B
C
H
P
I
K
L

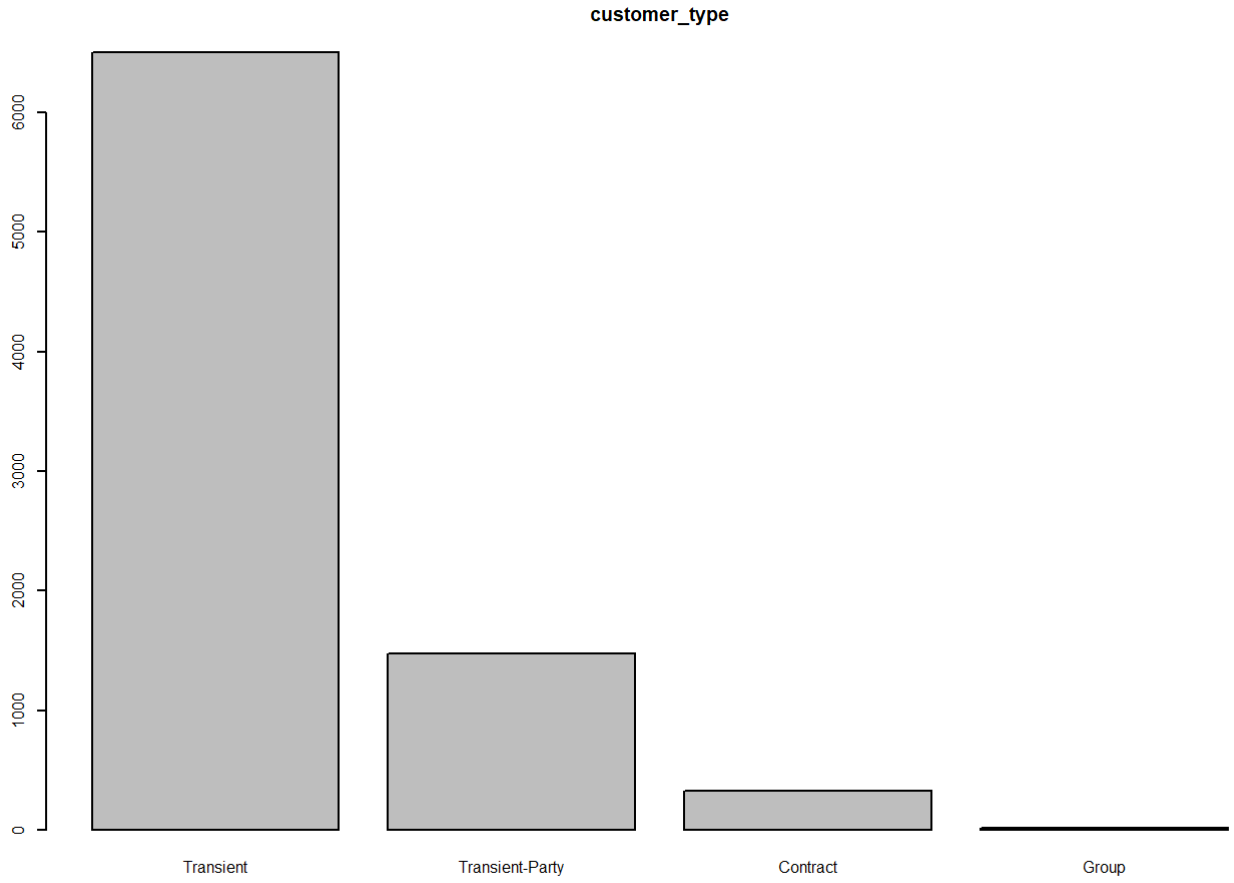

booking_changes
0
2000
4000
6000
0
1
2
3
4
6
7
8

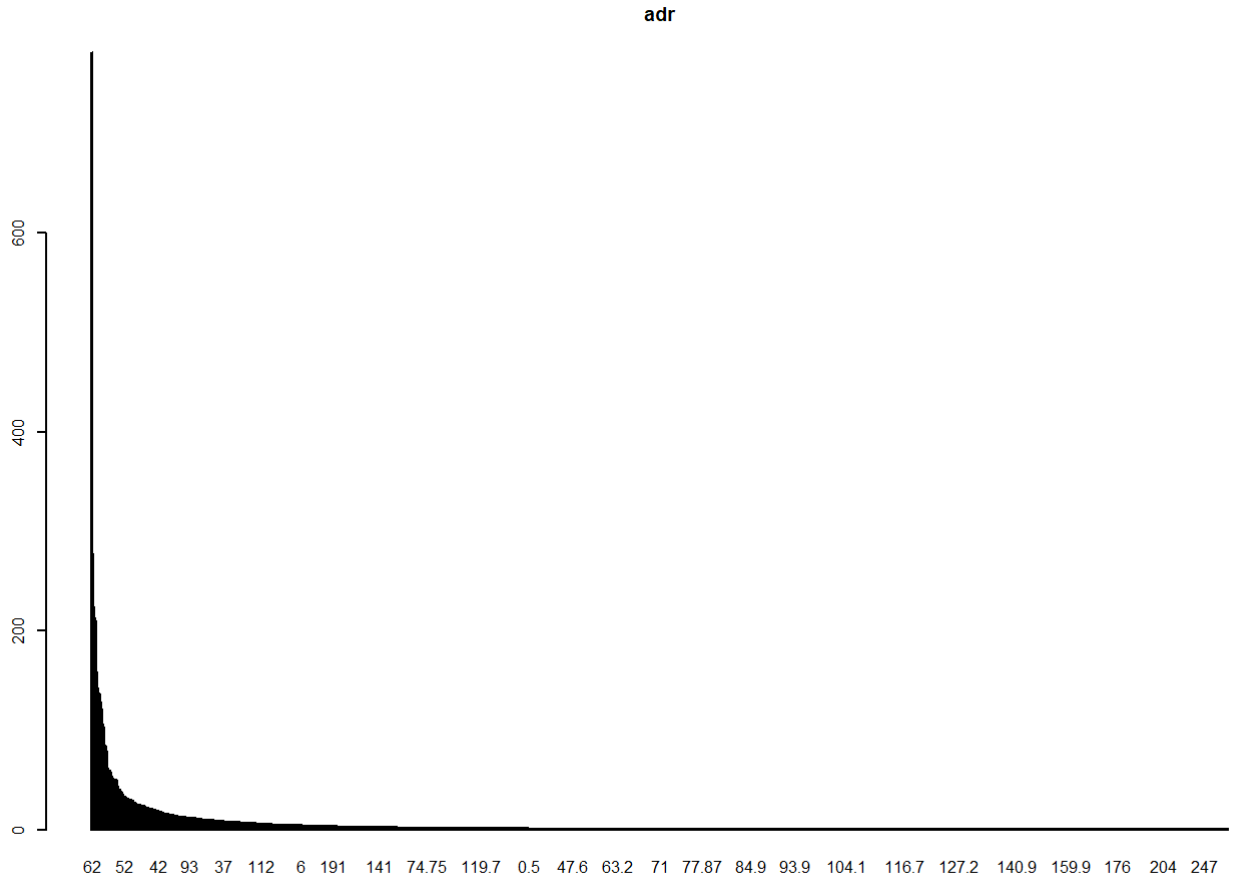

deposit_type
5000
4000
3000
2000
1000
0
No Deposit
Non Refund
Refundable

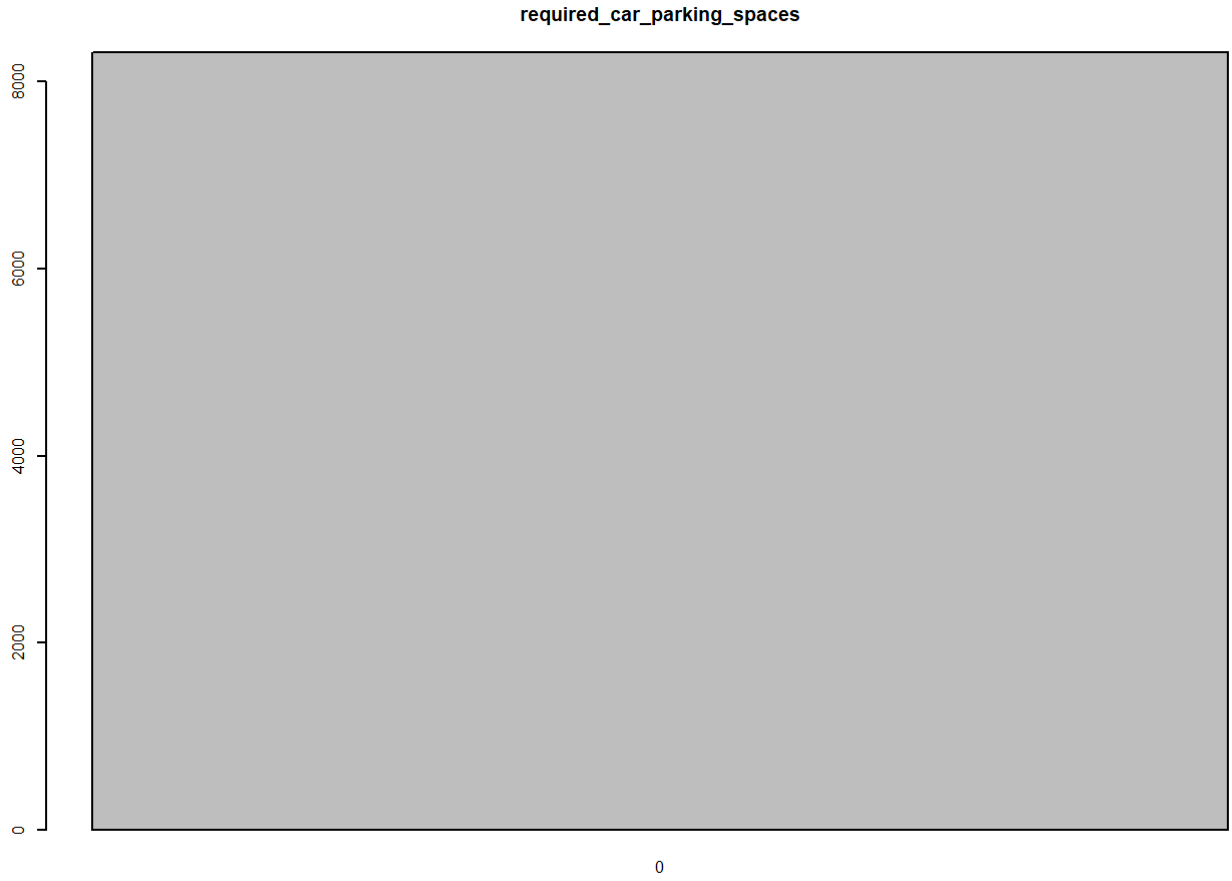

agent
2000
1500
1000
500
0
9 6 3 7 8 20 241 68 31 208 15 281 147 196 175 82 98 78 94 368 95 52 143 185 330 66 159 244 355 77

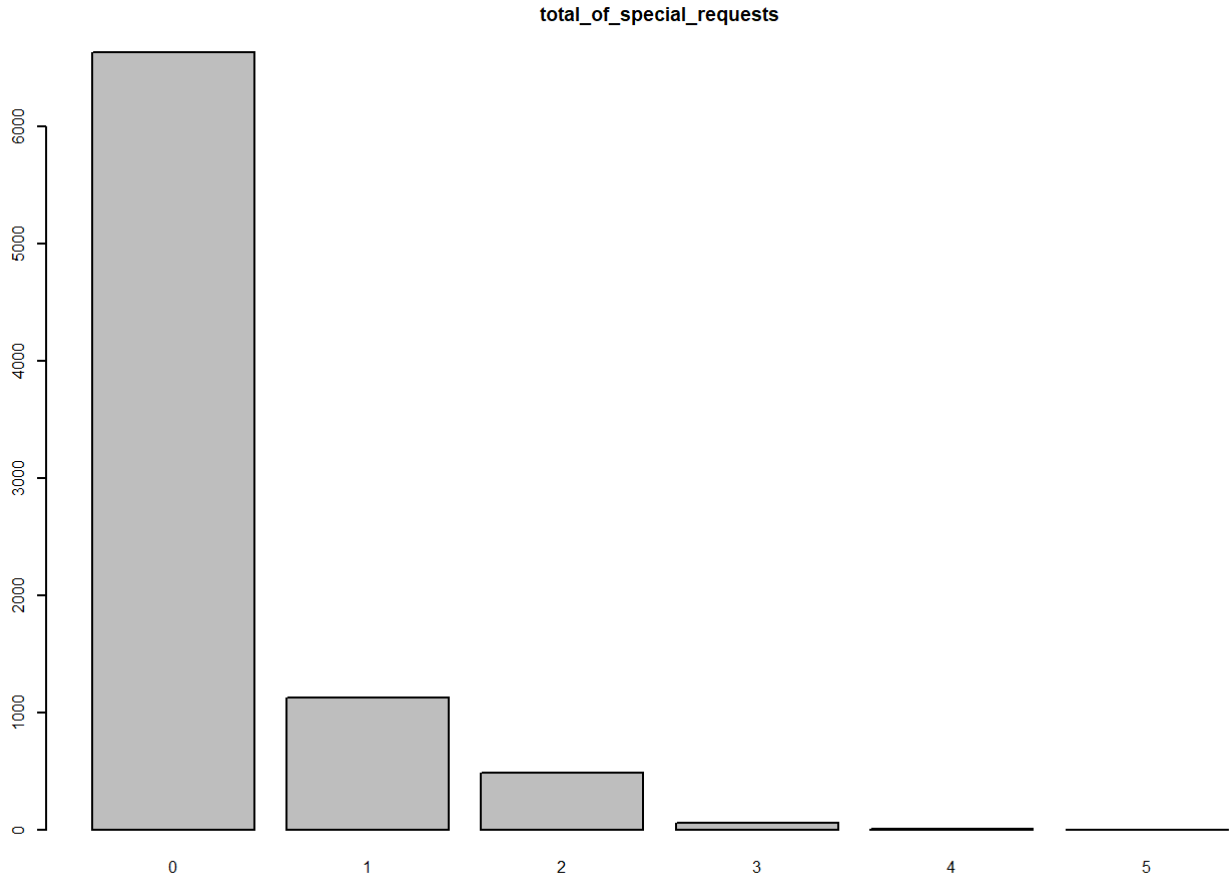

company
8000
6000
4000
2000
0
NULL 40 51 281 110 174 270 47 154 204 37 83 14 178 219 275 316 365 373 390 421 477 62 73 86


days_in_waiting_list
6000
4000
2000
0
0 31 44 28 3 17 391 21 120 19 42 9 48 79 15 80 34 160 16 68 4 76 1 8 55 107 207 236


customer_type
0
1000
2000
3000
4000
5000
6000
Transient
Transient-Party
Contract
Group


adr
0
200
400
600
62 52 42 93 37 112 6 191 141 74.75 119.7 0.5 47.6 63.2 71 77.87 84.9 93.9 104.1 116.7 127.2 140.9 159.9 176 204 247


required_car_parking_spaces
8000
6000
4000
2000
0
0


total_of_special_requests
6000
5000
4000
3000
2000
1000
0
0
1
2
3
4
5