NYC EDA

Hannah Khuong

Contents


```r
library(tidyverse)
library(skimr)
library(DataExplorer)
library(Hmisc)
library(gridExtra) # organize ggplot
library(lubridate) # time data
library(GGally) # quick eda plot
library(RColorBrewer)
library(zoo) # Fill NA
library(corrplot) # corr plot
library(kableExtra) # make table
```

```r
nyc <- read.csv("BOOKINGS_NYCHA.csv")
summary(nyc)
```

```
##                     product_type         stay_dt            dow
##  BUSINESS TRAVEL AGENCIES: 5429   02/01/2012:  671   Min.   :1.000
##  CORPORATE               : 5429   02/02/2012:  671   1st Qu.:2.000
##  FENCED                  : 5429   02/03/2012:  671   Median :4.000
##  GROUP                   : 5429   02/04/2012:  671   Mean   :4.043
##  MEMBERSHIP MARKETING    : 5429   02/07/2012:  671   3rd Qu.:6.000
##  OPAQUE                  : 5429   02/08/2012:  671   Max.   :7.000
##  (Other)                 :25864   (Other)   :54412
##       booking_dt      days_prior daily_gross_bookings daily_gross_rev
##  02/14/2012:  661   Min.   : 0   Min.   :  0.000      Min.   :    0.0
##  02/15/2012:  661   1st Qu.:15   1st Qu.:  0.000      1st Qu.:    0.0
##  02/16/2012:  661   Median :30   Median :  0.000      Median :    0.0
##  02/17/2012:  661   Mean   :30   Mean   :  1.621      Mean   :  406.9
##  02/18/2012:  661   3rd Qu.:45   3rd Qu.:  1.000      3rd Qu.:  255.0
##  02/19/2012:  661   Max.   :60   Max.   :257.000      Max.   :66315.0
##  (Other)   :54472
##  daily_cxl_bookings daily_cxl_rev      daily_net_bookings
##  Min.   : 0.0000    Min.   :     0.00   Min.   :-26.000
```

```
##  1st Qu.: 0.0000    1st Qu.:    0.00   1st Qu.:  0.000
##  Median : 0.0000    Median :    0.00   Median :  0.000
##  Mean   : 0.2852    Mean   :   74.51   Mean   :  1.336
##  3rd Qu.: 0.0000    3rd Qu.:    0.00   3rd Qu.:  1.000
##  Max.   :81.0000    Max.   :16119.00   Max.   :225.000
##
##  daily_net_rev     cummulative_gross_bookings cummulative_gross_rev
##  Min.   :-5961.0   Min.   :  0.00             Min.   :     0
##  1st Qu.:    0.0   1st Qu.:  6.00             1st Qu.:  1461
##  Median :    0.0   Median : 25.00             Median :  5749
##  Mean   :  332.4   Mean   : 61.59             Mean   : 15170
##  3rd Qu.:  175.0   3rd Qu.: 75.00             3rd Qu.: 17458
##  Max.   :66056.0   Max.   :835.00             Max.   :266552
##
##  cummulative_cxl_bookings cummulative_cxl_rev      OTB
##  Min.   :  0.000          Min.   :    0.0     Min.   :  0.00
##  1st Qu.:  0.000          1st Qu.:    0.0     1st Qu.:  5.00
##  Median :  2.000          Median :  418.5     Median : 22.00
##  Mean   :  6.557          Mean   : 1690.7     Mean   : 55.03
##  3rd Qu.:  7.000          3rd Qu.: 1790.2     3rd Qu.: 67.00
##  Max.   :317.000          Max.   :63273.0     Max.   :783.00
##
##     OTB_rev       OTB_to_be_cxl    OTB_rev_to_be_cxl OTB_to_survive
##  Min.   :     0   Min.   : 0.000   Min.   :    0     Min.   :  0.00
##  1st Qu.:  1257   1st Qu.: 0.000   1st Qu.:    0     1st Qu.:  5.00
##  Median :  5033   Median : 2.000   Median :  318     Median : 19.00
##  Mean   : 13480   Mean   : 3.845   Mean   :  988     Mean   : 51.19
##  3rd Qu.: 15693   3rd Qu.: 5.000   3rd Qu.: 1114     3rd Qu.: 62.00
##  Max.   :235930   Max.   :67.000   Max.   :21419     Max.   :770.00
##
##  OTB_rev_to_survive
##  Min.   :     0
##  1st Qu.:  1016
##  Median :  4581
##  Mean   : 12492
##  3rd Qu.: 14298
##  Max.   :224901
##
```

Data manipulation

Add variables

Add levels for product type

```r
# Make new column
nyc <- cbind(product_type_2 = 'Other', nyc)
nyc <- cbind(product_type_1 = 'Individual Transient', nyc)

# create product type group
closed_offer <- c('MEMBERSHIP MARKETING', 'TACTICAL MARKETING')
managed_business <- c('CORPORATE', 'GOVERNMENT', 'WHOLESALE', 'BUSINESS TRAVEL AGENCIES')
```

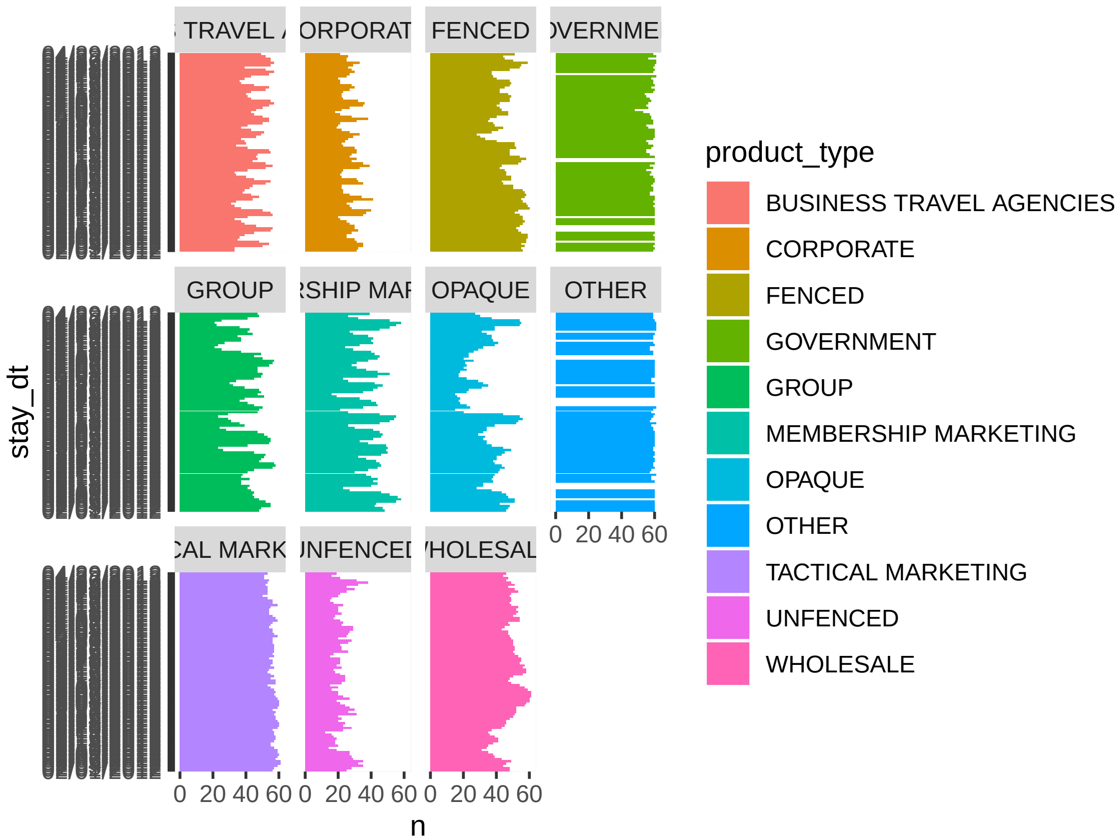
```
other <- c('OPAQUE', 'OTHER')
public_retail <- c('FENCED', 'UNFENCED')

# Rename vars in product type level 2
nyc$product_type_2 <- ifelse(nyc$product_type %in% closed_offer, 'Closed Offer',
                             ifelse(nyc$product_type %in% managed_business, 'Managed Business',
                                    ifelse(nyc$product_type %in% other, 'Other',
                                           ifelse(nyc$product_type %in% public_retail, 'Public Retail',
                                                  ifelse(nyc$product_type == 'GROUP', 'Group', 'Other')
# Rename vars in product type level 1 (biggest)
nyc$product_type_1 <- ifelse(nyc$product_type_2 == 'Group', 'Group','Individual Transient')
```

Calculate room price (by days prior, product type, and stay date)

These bars represent rows that we will roll-fill NA in

- Missing value = 60 means that `product_type` for that `stay_date` does not have any booking throughout 60-days booking window.
- The number of rows we lose due to no-booking after filling NA is about 1000 rows (about 16 `stay_dt`)
- Some does not have any booking initially (further away from `stay_dt`). This is why we have to backward fill then forward fill

```
# Examine NA situation
nyc %>%
  mutate(room_price = daily_gross_rev / daily_gross_bookings ) %>%
  filter(is.na(room_price)) %>%
  group_by(product_type, stay_dt, days_prior) %>%
  count(value = is.na(room_price)) %>%
  ggplot(aes(y = n, x = stay_dt, fill = product_type)) +
  geom_bar(stat = 'identity') + coord_flip() + facet_wrap(~product_type)
```




```
#fill na by roll over
room_price <- nyc %>%
  group_by(product_type, stay_dt, days_prior) %>%
  summarise(room_price = daily_gross_rev / daily_gross_bookings) %>%
  do(na.locf(., na.rm = FALSE, fromLast = TRUE)) %>% #roll backward first
  do(na.locf(., na.rm = FALSE)) # then roll backward (to fill in initial booking dates)

# Add room price column
nyc <-data.frame(nyc, room_price[4])
```

Rename DOW

```
nyc <- nyc %>%
   mutate(dow = case_when(dow == 1  ~ "Sun",
                          dow == 2 ~ "Mon",
                          dow == 3 ~ "Tue",
                          dow == 4 ~ 'Wed',
                          dow == 5 ~ 'Thu',
                          dow == 6 ~ 'Fri',
                          dow == 7 ~ 'Sat'))

# Establish order for dow
nyc$dow <- factor(nyc$dow,
                  levels = c('Sun', 'Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sat'))
```

Split Train Test

```r
# Convert to date
nyc$stay_dt <- as.Date(nyc$stay_dt, c('%m/%d/%Y'))
nyc$booking_dt <- as.Date(nyc$booking_dt, c('%m/%d/%Y'))

# Training data till 04/08/2012
train <- subset(nyc, stay_dt < as.Date("2012-04-09") )
# Testing data from 04/09/2012 - 04/29/2012
test <- subset(nyc, stay_dt > as.Date("2012-04-08") )
```

Calculate cancellation rate (train set only)

This is our target prediction variable. It is the `number of cancellation to come` (retrospectively calculated) over `On The Book (OTB) bookings` (cummulative net bookings).

```r
train <- train %>%
  mutate(cxl_rate = OTB_to_be_cxl / OTB) %>%
  mutate(cxl_rate = ifelse(is.na(cxl_rate),0,cxl_rate)) # When OTB = 0, rate = NA

# Check cxl_rate stats
summary(train$cxl_rate)
```

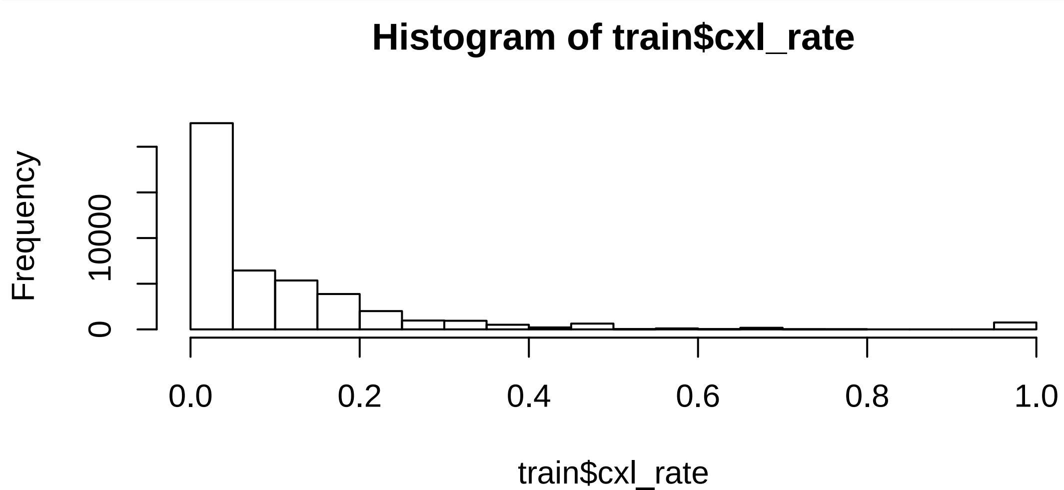
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.04902 0.10439 0.14286 1.00000
```

EDA

Univariate EDA

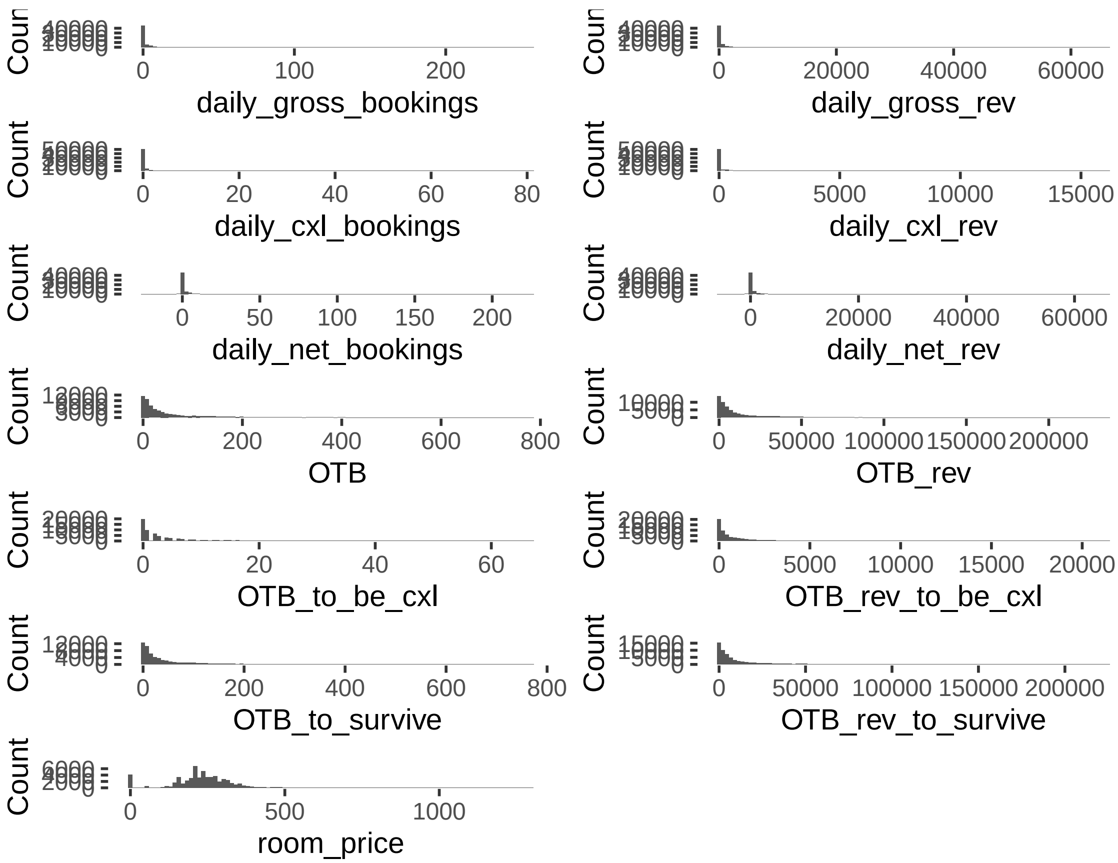
Dependent Var - Cancellation Rate

```r
hist(train$cxl_rate)
```



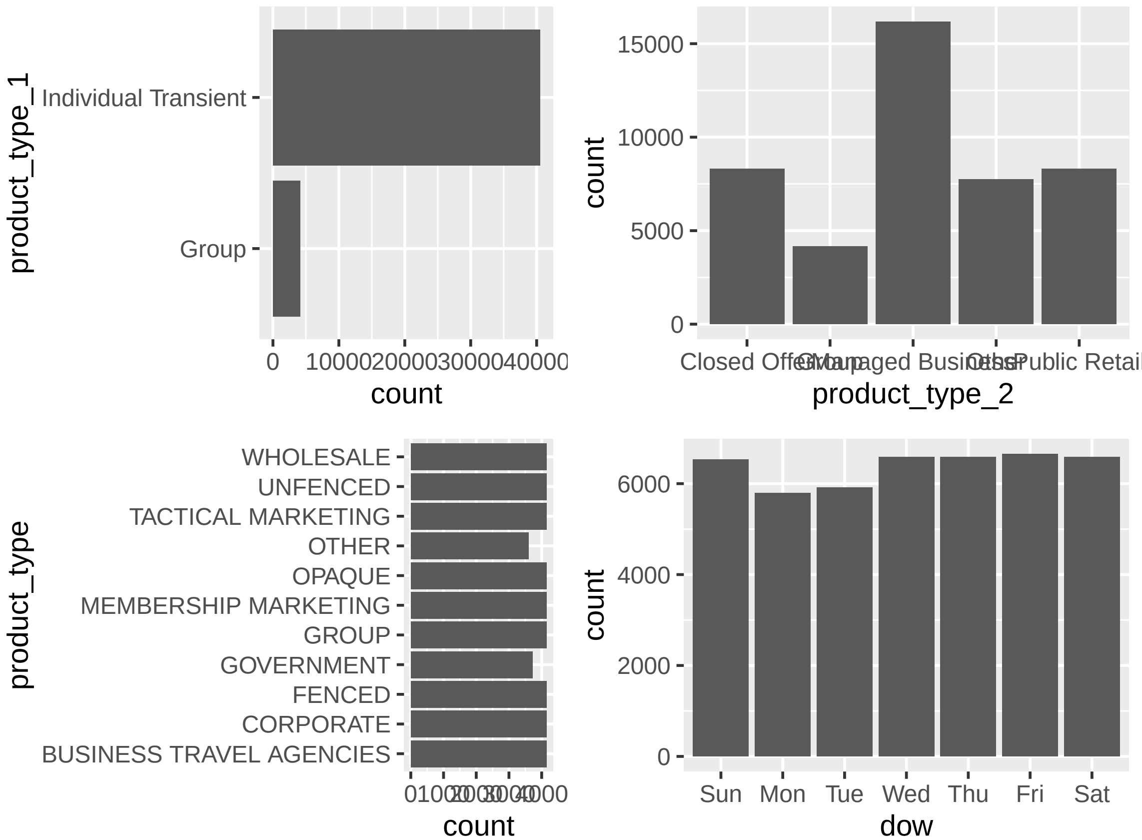
Continuous Variables

```r
#Count histogram
count_hist<- function(x){
  x + geom_histogram(bins = 100)+
    theme_bw() +
    theme(panel.border = element_blank(),
                       panel.grid.major = element_blank(),
                       panel.grid.minor = element_blank()) +
    labs(y ="Count")
}

# Create bar graphs
grid.arrange(
count_hist(ggplot(nyc, aes(daily_gross_bookings))),
count_hist(ggplot(nyc, aes(daily_gross_rev))),
count_hist(ggplot(nyc, aes(daily_cxl_bookings))),
count_hist(ggplot(nyc, aes(daily_cxl_rev))),
count_hist(ggplot(nyc, aes(daily_net_bookings))),
count_hist(ggplot(nyc, aes(daily_net_rev))),
count_hist(ggplot(nyc, aes(OTB))),
count_hist(ggplot(nyc, aes(OTB_rev))),
count_hist(ggplot(nyc, aes(OTB_to_be_cxl))),
count_hist(ggplot(nyc, aes(OTB_rev_to_be_cxl))),
count_hist(ggplot(nyc, aes(OTB_to_survive))),
count_hist(ggplot(nyc, aes(OTB_rev_to_survive))),
count_hist(ggplot(nyc, aes(room_price))), ncol = 2)
```



Categorical Variables

```r
# Create bar charts
grid.arrange(
  train %>% ggplot(aes(product_type_1))+ geom_bar() + coord_flip(),
  train %>% ggplot(aes(product_type_2))+ geom_bar(),
  train %>% ggplot(aes(product_type))+coord_flip()+geom_bar(),
  train %>% ggplot(aes(dow))+geom_bar(),
  ncol = 2)
```

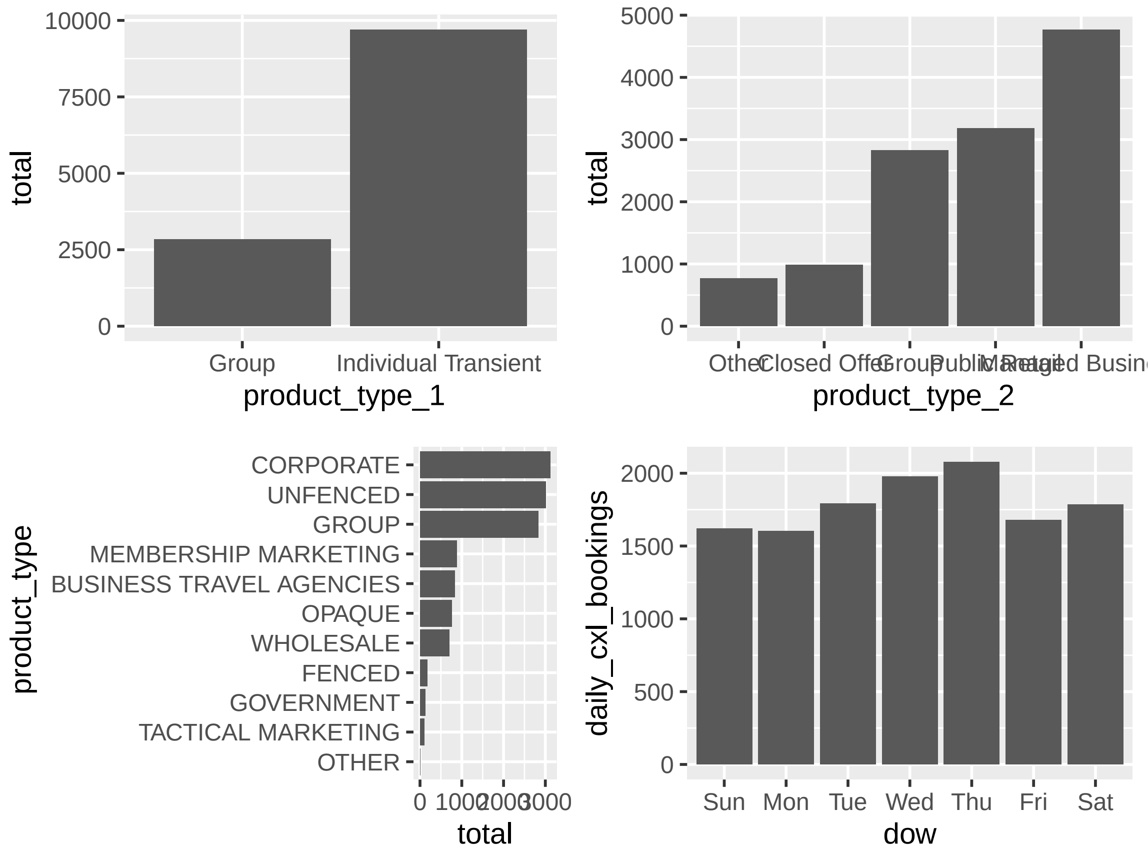


Multivariate EDA

```
# Create funtion
multi_bar_graph <- function(data, x, y){
  x <- enquo(x)
  y <-enquo(y)
  data %>%
    group_by(!!x) %>%
    select(!!x, !!y) %>%
    summarise(total = sum(!!y)) %>%
    ggplot(aes(x = reorder(!!x, total), y = total))+ geom_bar(stat = 'identity') + labs(x = x)
}
```

Cat. Var. By Number of Cancelled Bookings

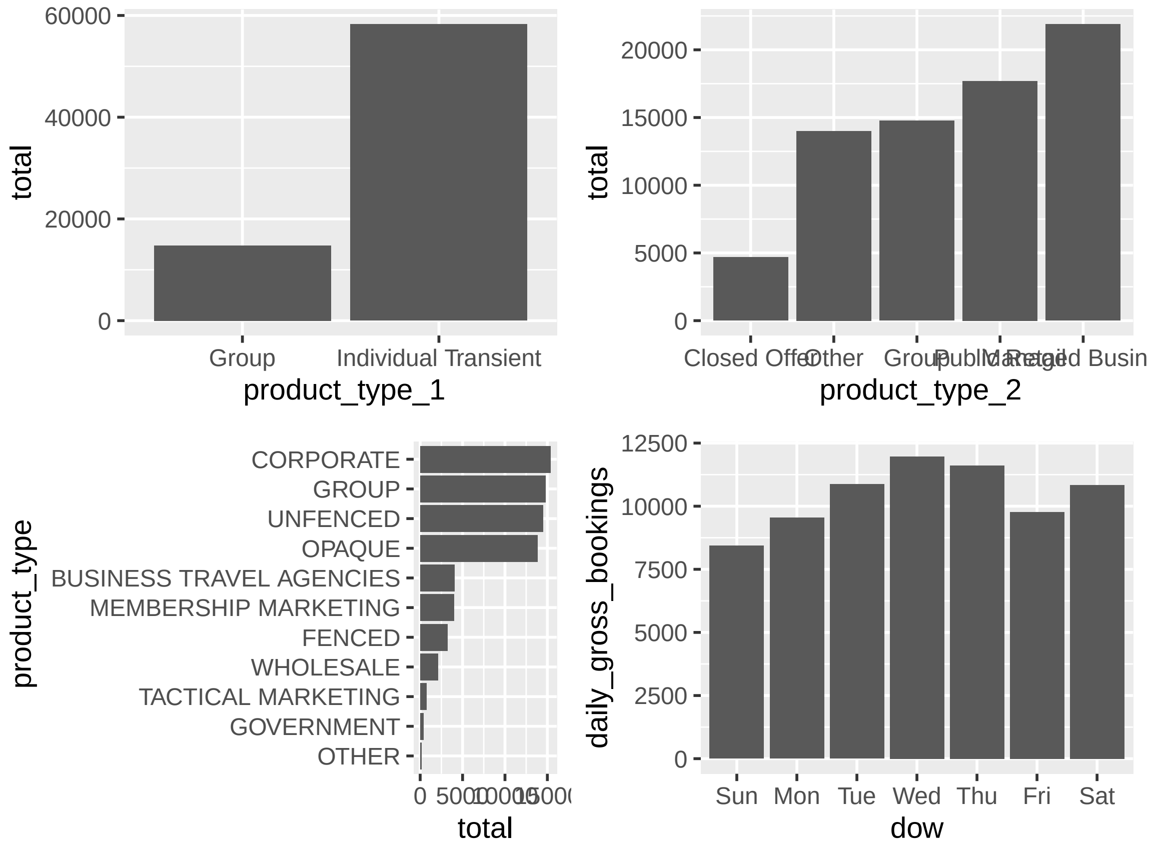
```
grid.arrange(
   multi_bar_graph(train, product_type_1, daily_cxl_bookings),
    multi_bar_graph(train, product_type_2, daily_cxl_bookings),
   multi_bar_graph(train, product_type, daily_cxl_bookings) +coord_flip(),
   train %>% ggplot(aes(x = dow, y = daily_cxl_bookings)) + stat_summary(fun.y = 'sum', geom = 'bar'),
   ncol = 2)
```



- Weekdays (especially Tue, Wed, Thu) have higher daily cxl bookings then weekends
- Top 3 product types with highest daily cxl bookings: corporate, unfenced, group

Cat. Var. By Daily Gross Bookings

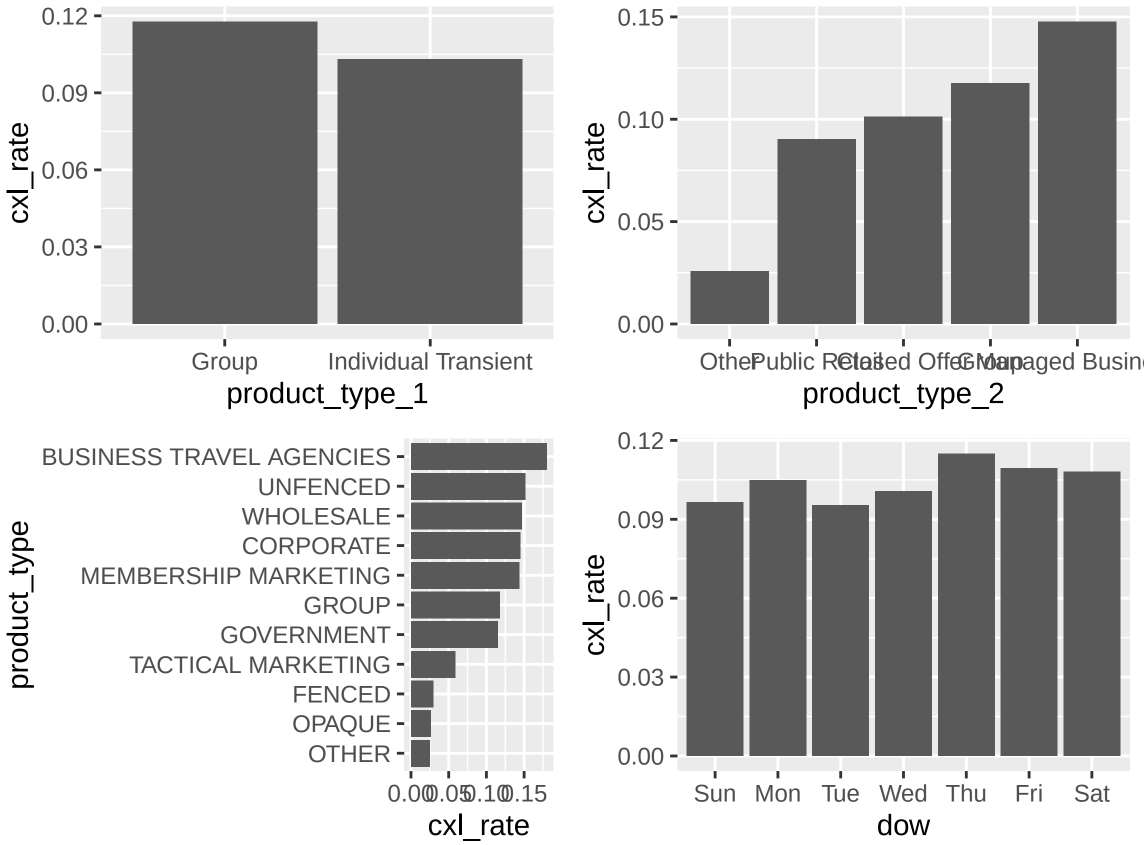
```
grid.arrange(
   multi_bar_graph(train, product_type_1, daily_gross_bookings),
    multi_bar_graph(train, product_type_2, daily_gross_bookings),
   multi_bar_graph(train, product_type, daily_gross_bookings) +coord_flip(),
   train %>% ggplot(aes(x = dow, y = daily_gross_bookings)) + stat_summary(fun.y = 'sum', geom = 'bar'),
   ncol = 2)
```



- Weekdays (especially Tue, Wed, Thu) have higher daily gross bookings amount than weekends
- Top 3 product types with highest daily gross bookings: corporate, group, unfenced

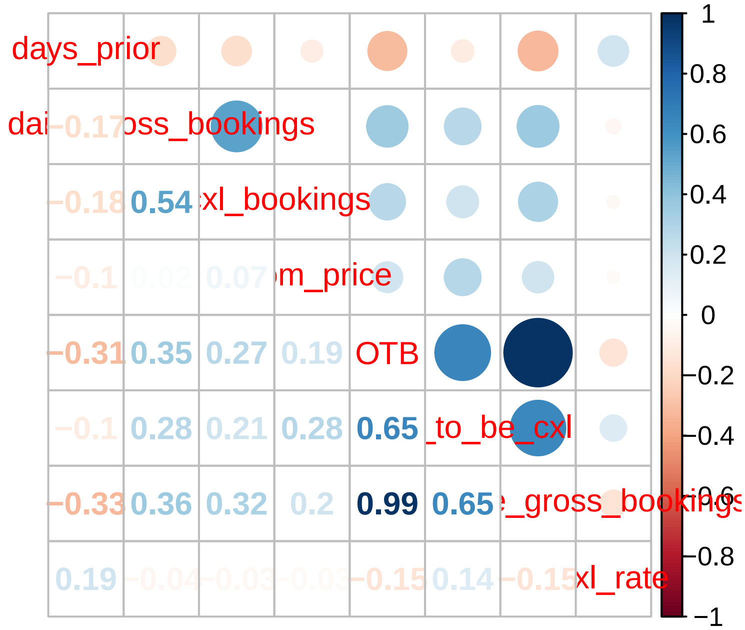
Cat. Var. By Cancellation Rate

```
grid.arrange(
  train %>% ggplot(aes(x = product_type_1, y = cxl_rate)) + stat_summary(fun.y = 'mean', geom = 'bar'),
  train %>% ggplot(aes(x = reorder(product_type_2, cxl_rate), y = cxl_rate)) + stat_summary(fun.y = 'mea
  train %>% ggplot(aes(x = reorder(product_type, cxl_rate), y = cxl_rate)) + stat_summary(fun.y = 'mean
  train %>% ggplot(aes(x = dow, y = cxl_rate)) + stat_summary(fun.y = 'mean', geom = 'bar'),
  ncol = 2)
```



Correlation Matrix of continuous variables

```
# Find correlation of quantitative variables
cor_plot <- train %>%
  filter(room_price > 1, na.omit(room_price)) %>%  #Filter promotion room_price and missing value in ro
  select(days_prior, daily_gross_bookings,daily_cxl_bookings,room_price, OTB,  OTB_to_be_cxl, cummulati
a <- cor(cor_plot)

corrplot.mixed(a)
```



- Room price doens't have a strong correlation with other factors (can ignore the impact of room price)

By days prior

Total trend

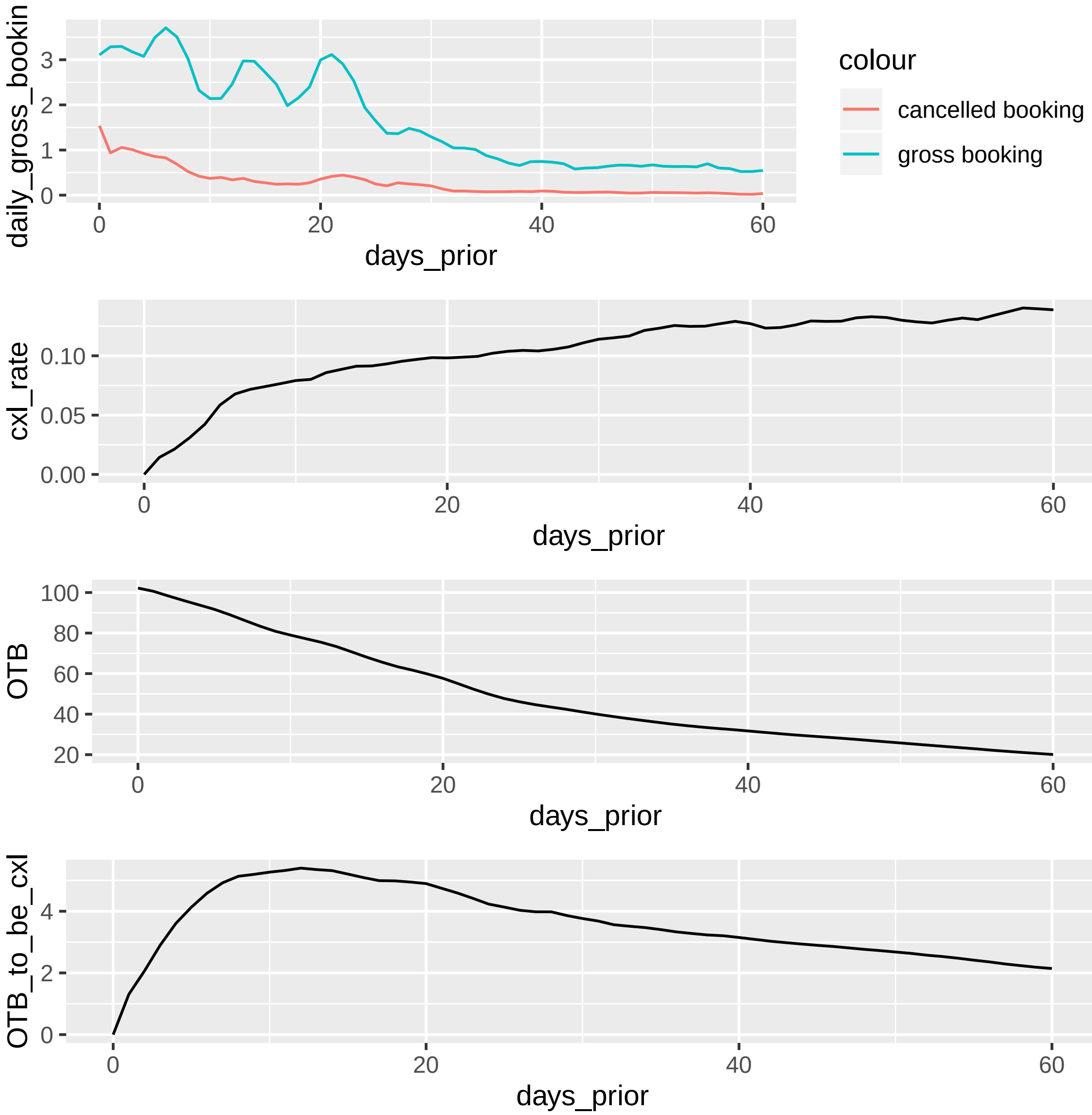
Graph 1: Cancellation and Booking both increase closer to the stay date.

Graph 2: Cancellation rate (to-be-cancelled / OTB) decreases approaching stay date because number of to-be-cancelled of OTB decreases

Graph 3: Number of OTB increases because this is cummulative value

Graph 4: OTB to be cancelled reach a peak at days prior 10

```
grid.arrange(
train %>%
  ggplot(aes(x = days_prior)) +
  stat_summary(aes(y = daily_gross_bookings, colour = 'gross booking'), fun.y = 'mean', geom = 'line') +
  stat_summary(aes(y = daily_cxl_bookings, colour = 'cancelled booking'), fun.y = 'mean', geom = 'line')
train %>%
  ggplot(aes(x = days_prior)) +
  stat_summary(aes(y = cxl_rate), fun.y = 'mean', geom = 'line'),
train %>%
  ggplot(aes(x = days_prior)) +
  stat_summary(aes(y = OTB), fun.y = 'mean', geom = 'line'),
train %>%
  ggplot(aes(x = days_prior)) +
  stat_summary(aes(y = OTB_to_be_cxl), fun.y = 'mean', geom = 'line'),
ncol = 1)
```



Trend by product groups

Graph 1: Most of product-types'cancellaton rate decrease as days prior decrease. The cancellation rate of Government and Other doesn't follow this pattern

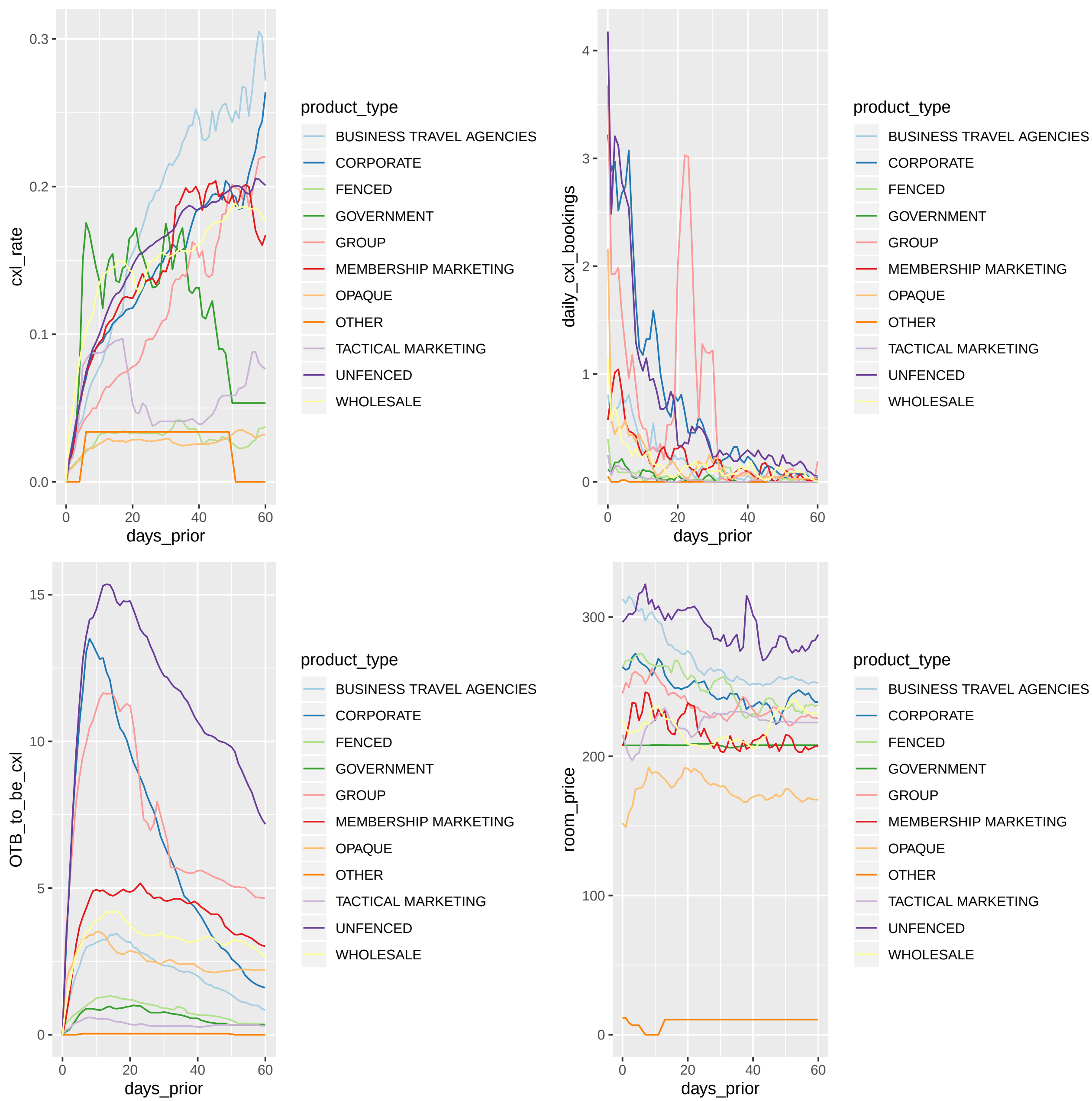
Graph 2: Abnormal pattern in Group product type, maybe caused by the cancellation of special events or mistake bookings.

Graph 3: OTB to be cxl of Unfenced, Corporate and Group are more volatile, peak reached in roughlt days prior 10. The value of Other, Tactical Marketing, Government and Fenced are more stable.

Graph 4: Room price of Unfenced group is highest.

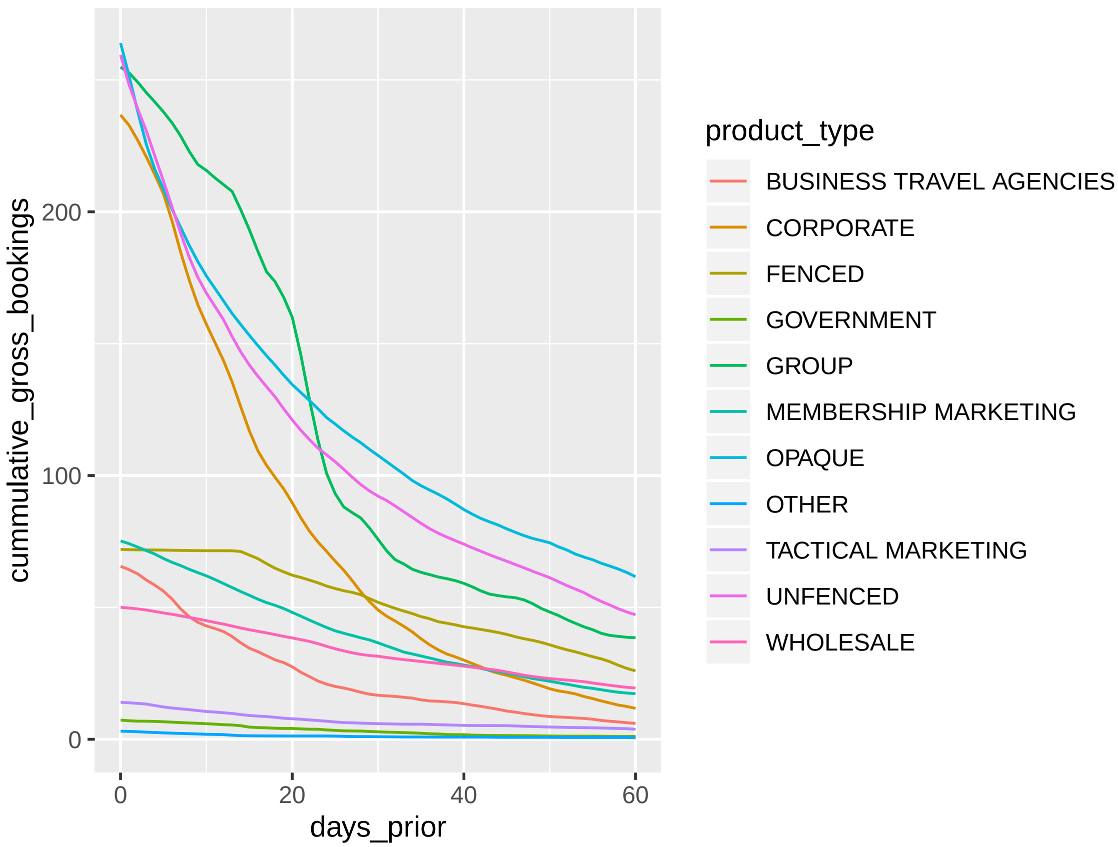
```
grid.arrange(
train %>%
  ggplot(aes(x = days_prior, color = product_type)) +
  stat_summary(aes(y = cxl_rate), fun.y = 'mean', geom = 'line')+
  scale_color_brewer(palette = "Paired"),
train %>%
```

```
  ggplot(aes(x = days_prior, color = product_type)) +
  stat_summary(aes(y = daily_cxl_bookings), fun.y = 'mean', geom = 'line')+
  scale_color_brewer(palette = "Paired"),
train %>%
  ggplot(aes(x = days_prior, color = product_type)) +
  stat_summary(aes(y = OTB_to_be_cxl), fun.y = 'mean', geom = 'line')+
  scale_color_brewer(palette = "Paired"),
train %>%
  ggplot(aes(x = days_prior, color = product_type)) +
  stat_summary(aes(y = room_price), fun.y = 'mean', geom = 'line')+
  scale_color_brewer(palette = "Paired"),
ncol = 2)
```



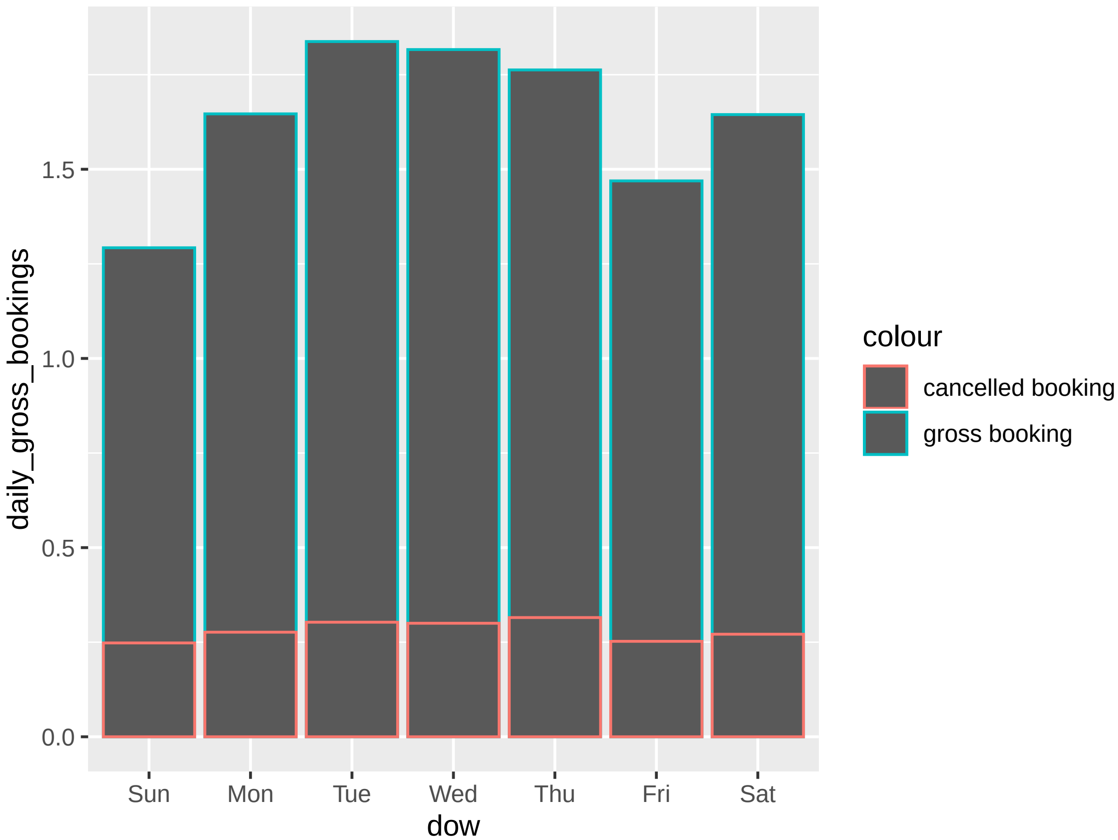
Cumulative gross bookings - demand level

```
train %>%
  ggplot(aes(x = days_prior, color = product_type)) +
  stat_summary(aes(y = cummulative_gross_bookings), fun.y = 'mean', geom = 'line')
```



By days of week

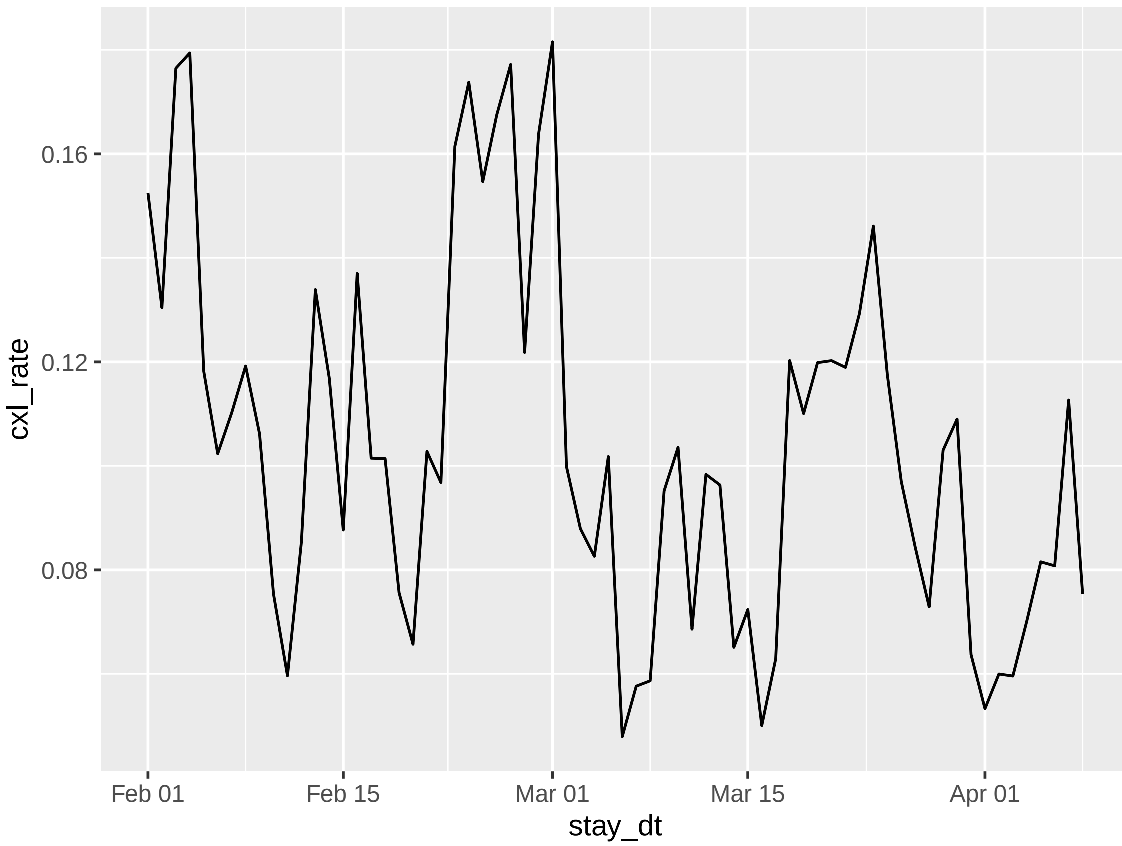
```
train %>%
  ggplot(aes(x = dow)) +
  stat_summary(aes(y = daily_gross_bookings, colour = 'gross booking'), fun.y = 'mean', geom = 'bar') +
  stat_summary(aes(y = daily_cxl_bookings, colour = 'cancelled booking'), fun.y = 'mean', geom = 'bar')
```



- Weekdays have higher booking amount and cxl booking amount.

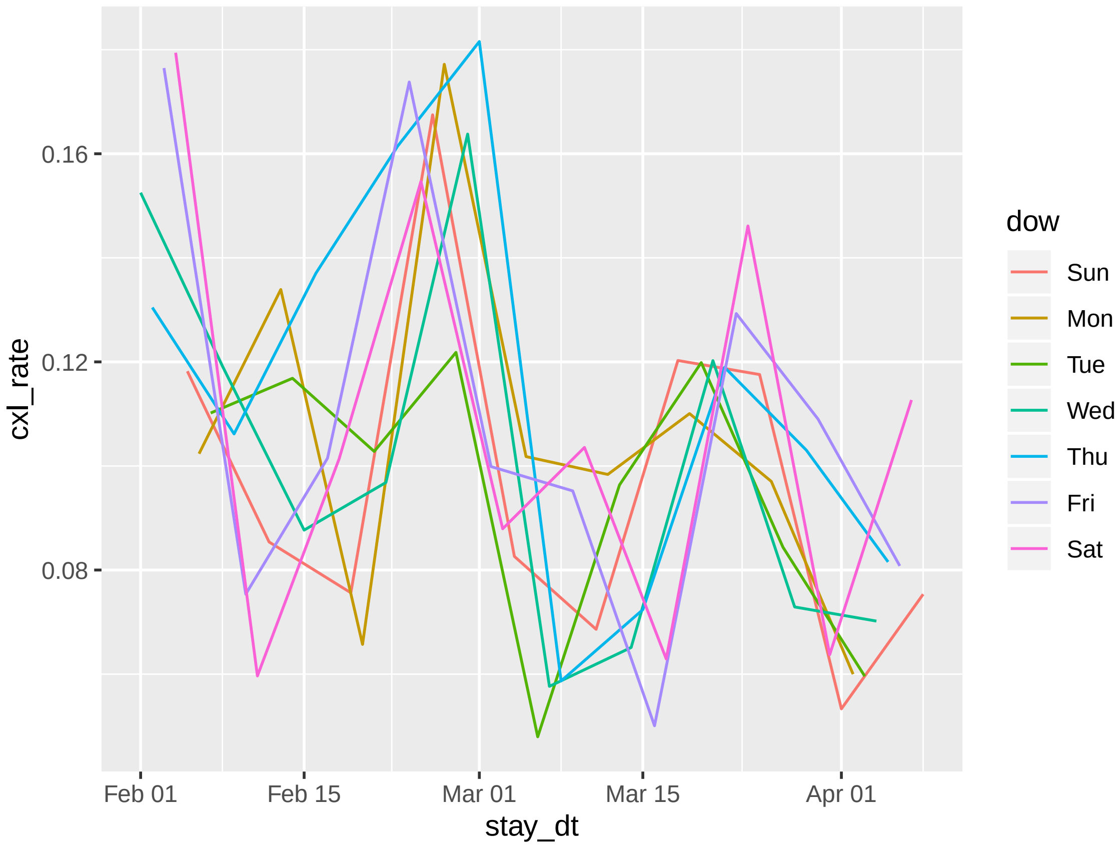
By stay date

```
train %>%
  ggplot(aes(x = stay_dt)) +
  stat_summary(aes(y = cxl_rate), fun.y = 'mean', geom = 'line')
```



There are 3 peaks of cancellation across `dow`: `Feb 01`, `Feb 20 - Mar 01` and `Mar 15 - Mar 16` There are 3 dips of cancallation across `dow`: `Feb 10`, `Mar 03`, and `Apr 01`

```
train %>%
  ggplot(aes(x = stay_dt, color = dow)) +
  stat_summary(aes(y = cxl_rate), fun.y = 'mean', geom = 'line')
```



Regrouping product type

Method 1: Regroup by cancellation rate trend with days prior

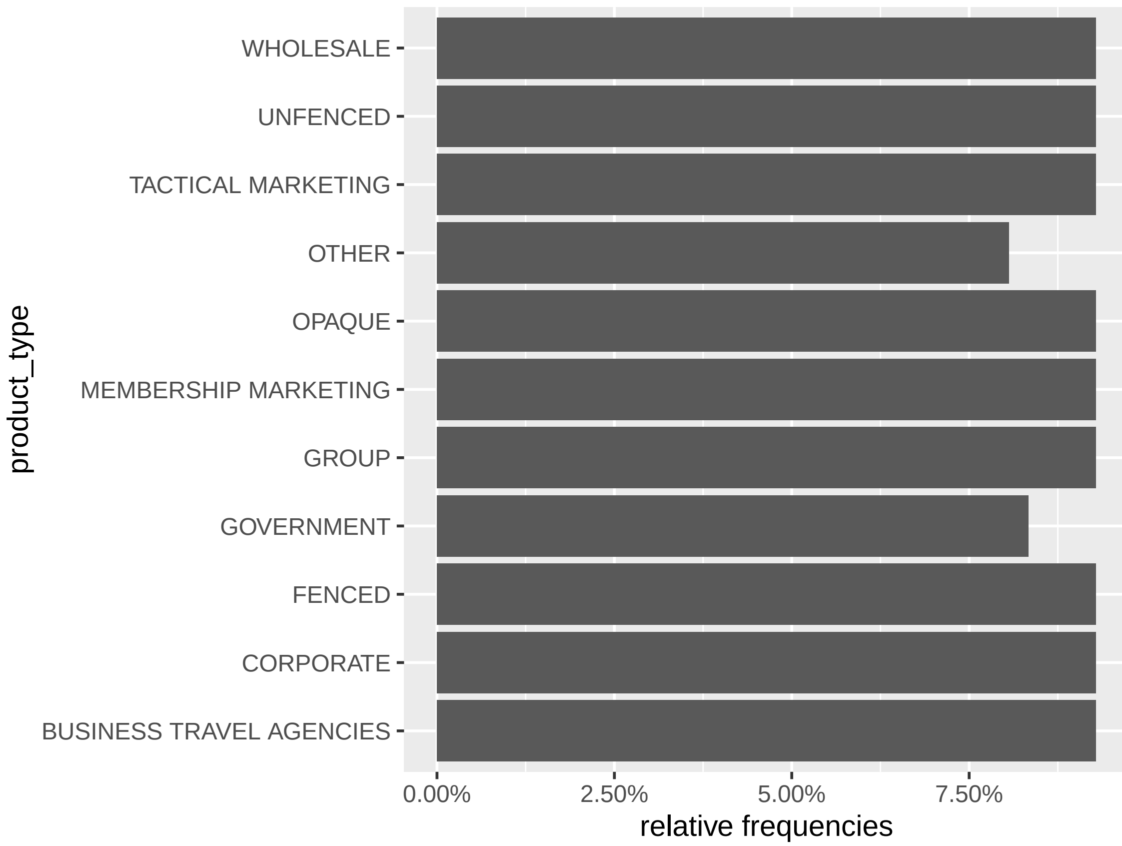
Number of cancellation of each product type:

```
kable(
train %>% group_by(product_type_2, product_type) %>% summarise(cancellation = sum(daily_cxl_bookings)))
  kable_styling(bootstrap_options = "striped", full_width = F)
```

product_type_2	product_type	cancellation
Closed Offer	MEMBERSHIP MARKETING	883
Closed Offer	TACTICAL MARKETING	100
Group	GROUP	2831
Managed Business	BUSINESS TRAVEL AGENCIES	836
Managed Business	CORPORATE	3113
Managed Business	GOVERNMENT	124
Managed Business	WHOLESALE	695
Other	OPAQUE	761
Other	OTHER	5
Public Retail	FENCED	170
Public Retail	UNFENCED	3012

Sample size

```r
train %>%
ggplot(aes(x = product_type)) +
  geom_bar(aes(y = (..count..)/sum(..count..)))+
  scale_y_continuous(labels=scales::percent) + coord_flip() +
  ylab("relative frequencies")
```



Cancellation trend of each product type. This is the main criteria for regrouping

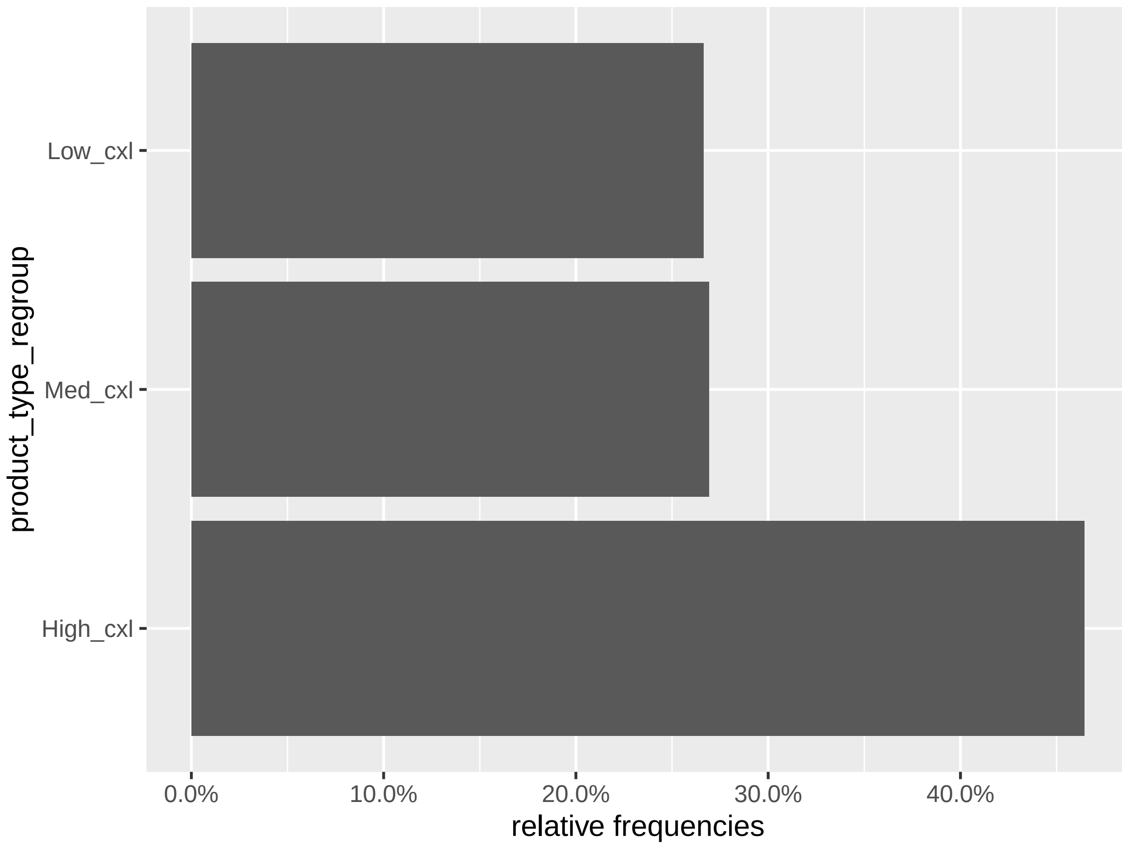
- High level of cancellation:
  - Business Travel Agencies (BTA)
  - Corporate
  - Unfenced
  - Membership marketing
  - Wholesale
- Middle level of cancellation
  - Group
  - Government
  - Tactical Marketing
- Low lever of cancellation
  - Fenced
  - Other
  - Opague

```r
train %>%
  ggplot(aes(x = days_prior, color = product_type)) +
  stat_summary(aes(y = cxl_rate), fun.y = 'mean', geom = 'line') +
  scale_color_brewer(palette = "Paired")
```


```r
## First Regrouping
# create product type group
high_cxl <- c('MEMBERSHIP MARKETING', 'WHOLESALE', 'BUSINESS TRAVEL AGENCIES', 'UNFENCED', 'CORPORATE')
mid_cxl <- c('GROUP', 'GOVERNMENT','TACTICAL MARKETING' )
low_cxl <- c('OPAQUE', 'OTHER', 'FENCED' )

# Make Column in nyc dataset
nyc <- cbind(product_type_regroup = 'Other', nyc)
# Rename vars in product type level 2
nyc$product_type_regroup <- ifelse(nyc$product_type %in% high_cxl, 'High_cxl',
                              ifelse(nyc$product_type %in% mid_cxl, 'Med_cxl',
                                     ifelse(nyc$product_type %in% low_cxl, 'Low_cxl','Other')))

# Make Column in train dataset
train <- cbind(product_type_regroup = 'Other', train)
# Rename vars in product type level 2
train$product_type_regroup <- ifelse(train$product_type %in% high_cxl, 'High_cxl',
                              ifelse(train$product_type %in% mid_cxl, 'Med_cxl',
                                     ifelse(train$product_type %in% low_cxl, 'Low_cxl','Other')))

# Establish order
train$product_type_regroup <- factor(train$product_type_regroup,
                  levels = c('High_cxl', 'Med_cxl', 'Low_cxl'))
nyc$product_type_regroup <- factor(nyc$product_type_regroup,
                  levels = c('High_cxl', 'Med_cxl', 'Low_cxl'))
```

Check sample size

```r
train %>%
ggplot(aes(x = product_type_regroup)) +
  geom_bar(aes(y = (..count..)/sum(..count..)))+
  scale_y_continuous(labels=scales::percent) + coord_flip() +
  ylab("relative frequencies")
```



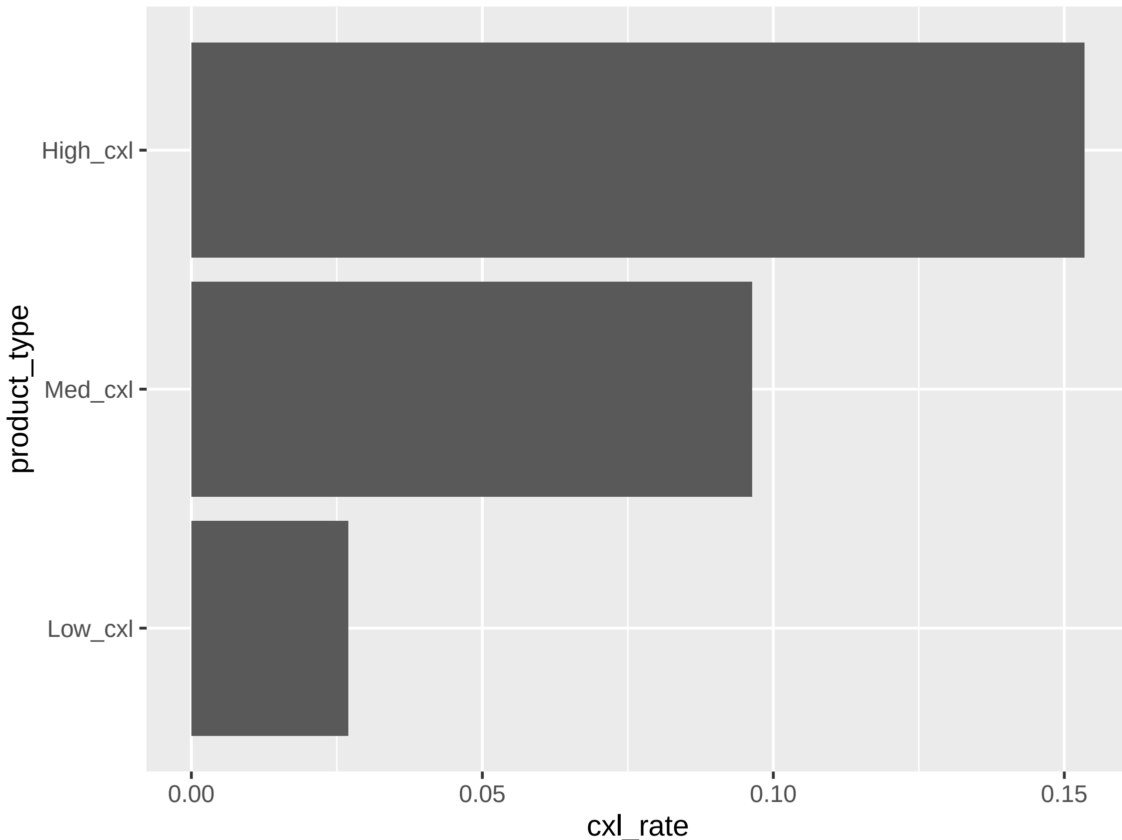
Check number of cancellations in each group

```r
kable(
train %>% group_by(product_type_regroup) %>% summarise(cancellation = sum(daily_cxl_bookings))) %>%
  kable_styling(bootstrap_options = "striped", full_width = F)
```

product_type_regroup	cancellation
High_cxl	8539
Med_cxl	3055
Low_cxl	936

EDA with new grouping

```r
train %>% ggplot(aes(x = reorder(product_type_regroup, cxl_rate), y = cxl_rate)) + stat_summary(fun.y =
```



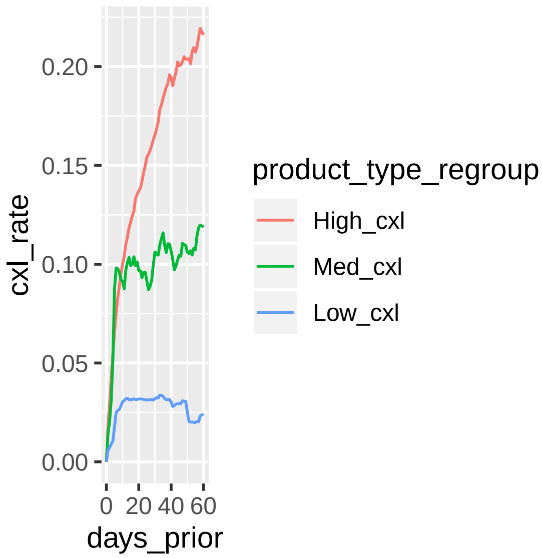
Relationship with days prior

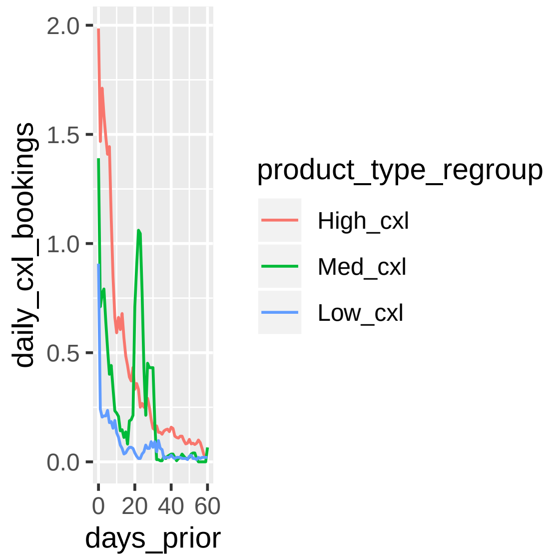
Cancellation rate trend vary greatly through days prior at 3 distinct levels

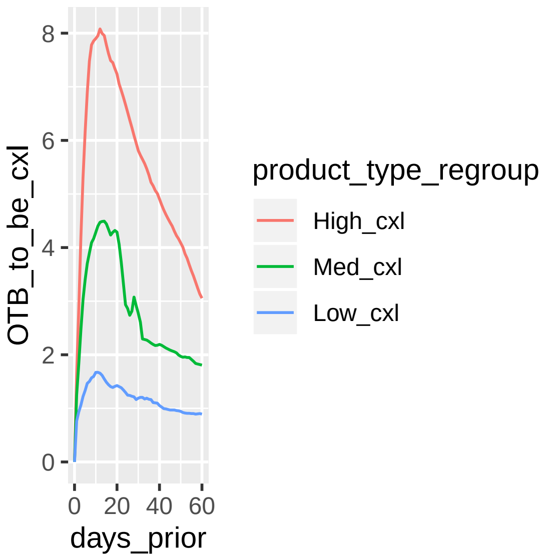
Daily cancellation for Med_cxl has a significant bump in 20-30 days priors

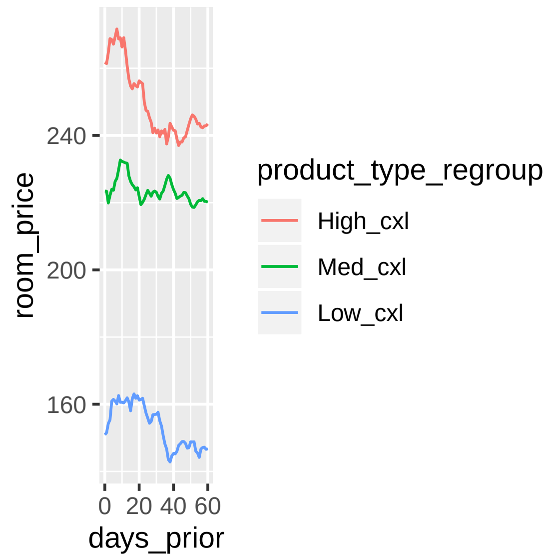
High and Low Cxl group has significantly higher room price than Low Cxl group

```
grid.arrange(
train %>%
  ggplot(aes(x = days_prior, color = product_type_regroup)) +
  stat_summary(aes(y = cxl_rate), fun.y = 'mean', geom = 'line'),
train %>%
  ggplot(aes(x = days_prior, color = product_type_regroup)) +
  stat_summary(aes(y = daily_cxl_bookings), fun.y = 'mean', geom = 'line'),
train %>%
  ggplot(aes(x = days_prior, color = product_type_regroup)) +
  stat_summary(aes(y = OTB_to_be_cxl), fun.y = 'mean', geom = 'line'),
train %>%
  ggplot(aes(x = days_prior, color = product_type_regroup)) +
  stat_summary(aes(y = room_price), fun.y = 'mean', geom = 'line'),
ncol = 2)
```



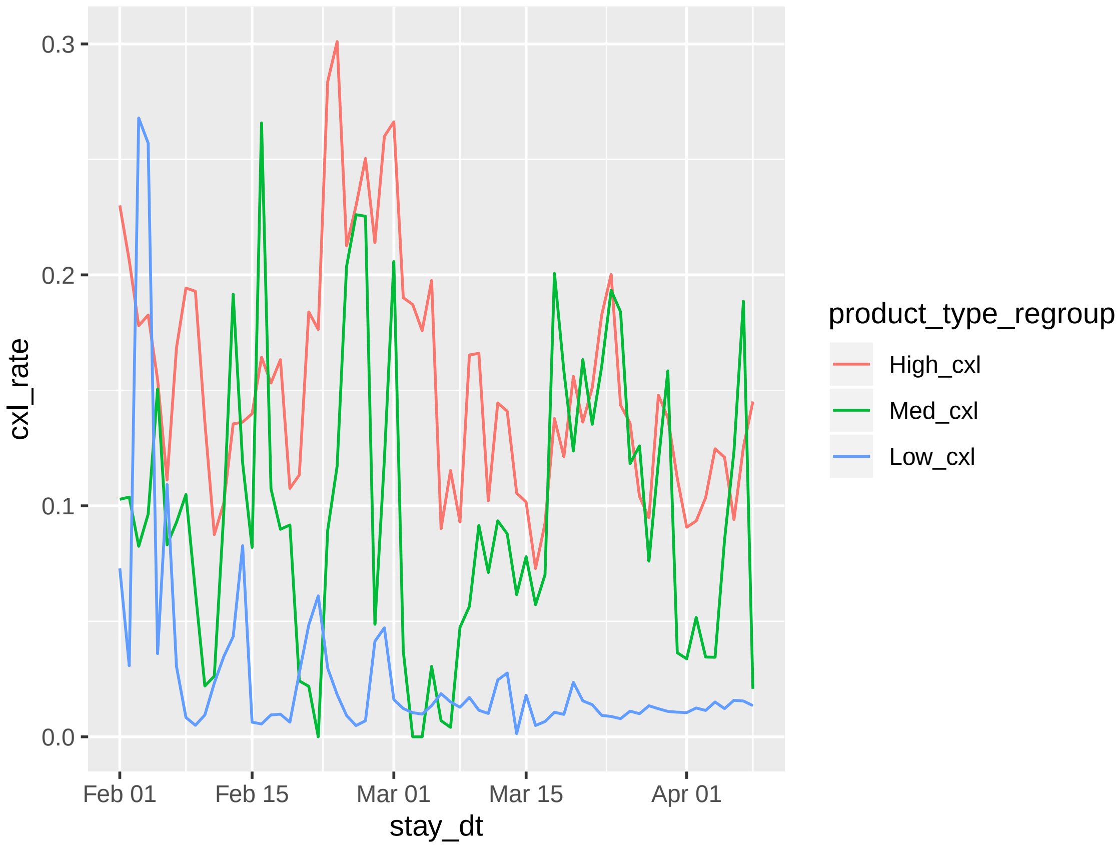






Relationship with stay_dt

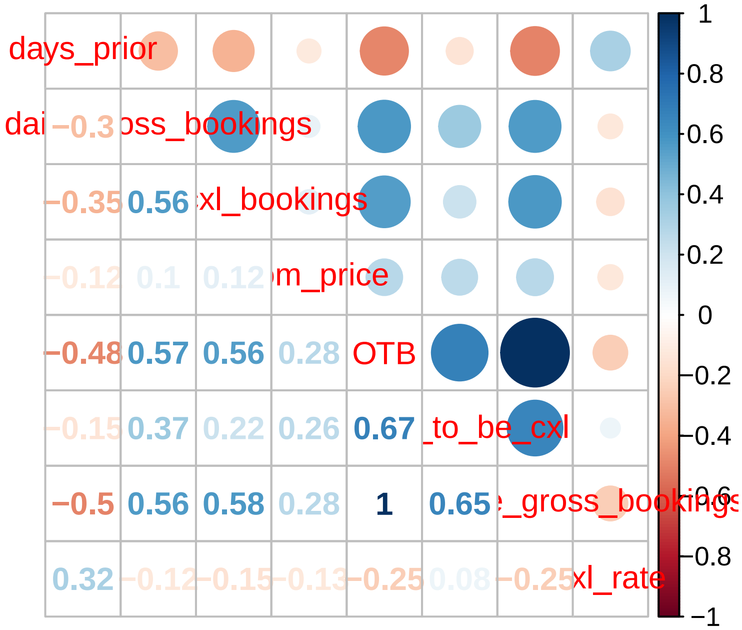
```
train %>%
  ggplot(aes(x = stay_dt, color = product_type_regroup)) +
  stat_summary(aes(y = cxl_rate), fun.y = 'mean', geom = 'line')
```



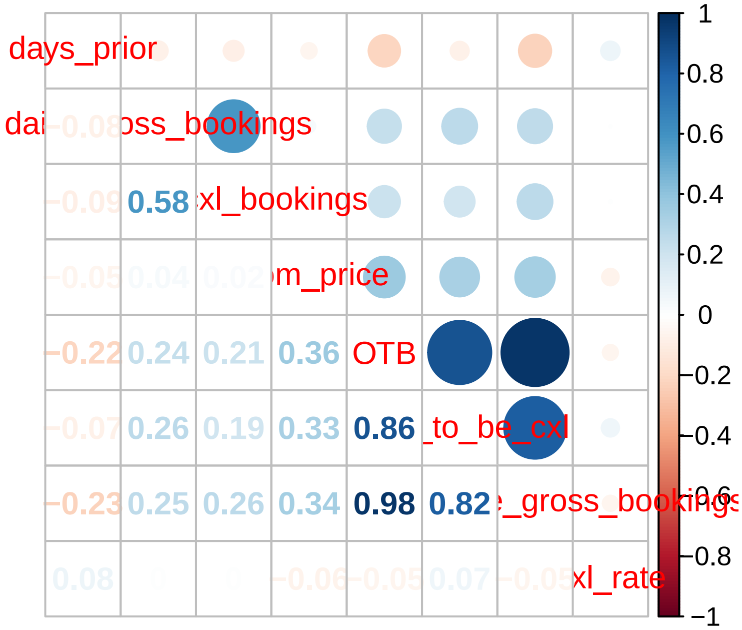
Correlational analysis

After regrouping, for different groups of product types, we can see the correlation change, which means that for different group, the impactor of cxl rate vary.

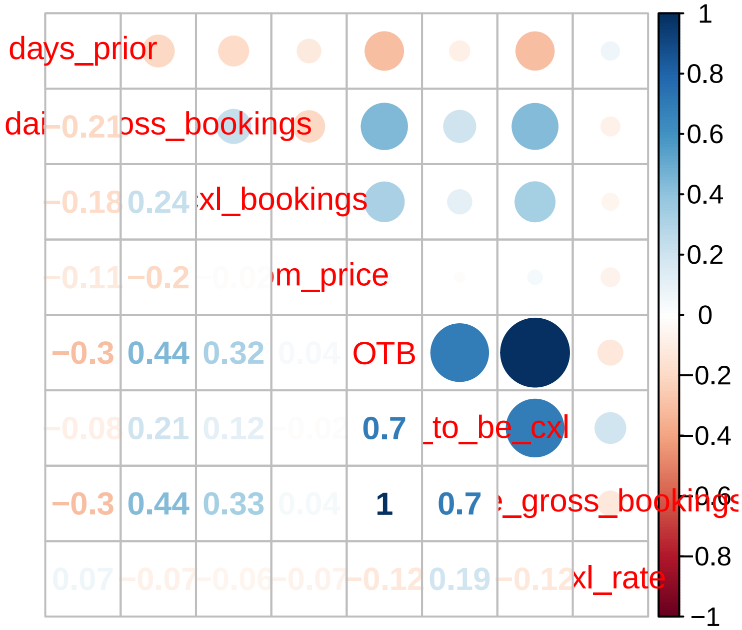
```
# Find correlation of quantitative variables

cor_plot <- train %>%
  filter(product_type_regroup == "High_cxl") %>%
  filter(room_price > 1, na.omit(room_price)) %>%  #Filter promotion room_price and missing value in ro
  select(days_prior, daily_gross_bookings,daily_cxl_bookings,room_price, OTB,  OTB_to_be_cxl, cummulati
a <- cor(cor_plot)

corrplot.mixed(a)
```




```r
cor_plot <- train %>%
  filter(product_type_regroup == "Med_cxl") %>%
  filter(room_price > 1, na.omit(room_price)) %>%  #Filter promotion room_price and missing value in ro
  select(days_prior, daily_gross_bookings,daily_cxl_bookings,room_price, OTB,  OTB_to_be_cxl, cummulati
a <- cor(cor_plot)
corrplot.mixed(a)
```



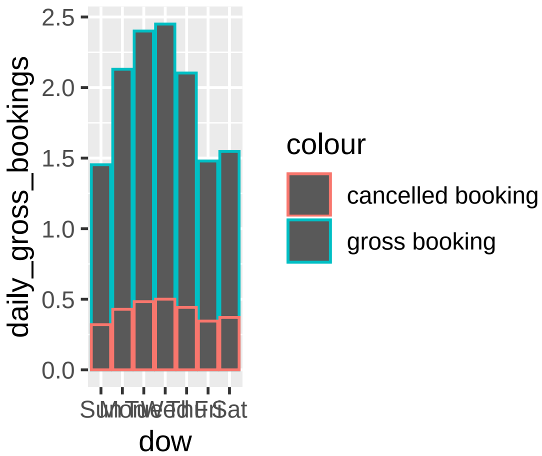

```r
cor_plot <- train %>%
  filter(product_type_regroup == "Low_cxl") %>%
  filter(room_price > 1, na.omit(room_price)) %>%  #Filter promotion room_price and missing value in ro
  select(days_prior, daily_gross_bookings,daily_cxl_bookings,room_price, OTB,  OTB_to_be_cxl, cummulati
a <- cor(cor_plot)
corrplot.mixed(a)
```

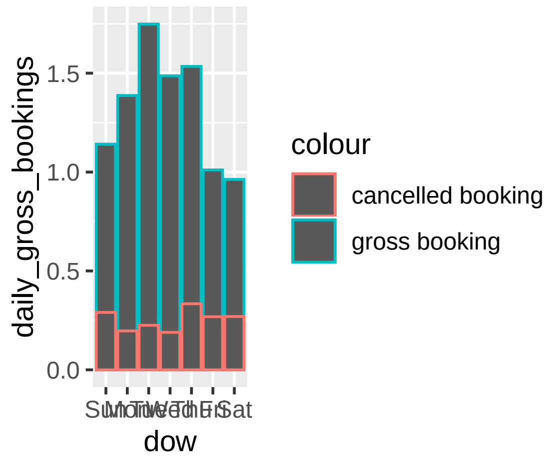


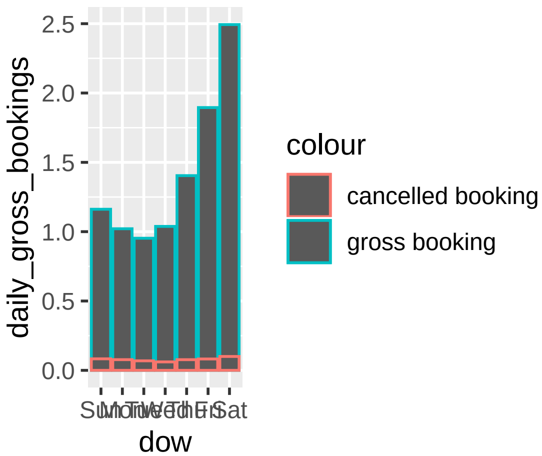
Relationship with DOW

We see that High and Medium Cxl group has similar pattern but Low Cxl group

```r
grid.arrange(

train %>%
  filter(product_type_regroup == "High_cxl") %>%
  ggplot(aes(x = dow)) +
  stat_summary(aes(y = daily_gross_bookings, colour = 'gross booking'), fun.y = 'mean', geom = 'bar') +
  stat_summary(aes(y = daily_cxl_bookings, colour = 'cancelled booking'), fun.y = 'mean', geom = 'bar')
train %>%
  filter(product_type_regroup == "Med_cxl") %>%
  ggplot(aes(x = dow)) +
  stat_summary(aes(y = daily_gross_bookings, colour = 'gross booking'), fun.y = 'mean', geom = 'bar') +
  stat_summary(aes(y = daily_cxl_bookings, colour = 'cancelled booking'), fun.y = 'mean', geom = 'bar')
train %>%
  filter(product_type_regroup == "Low_cxl") %>%
  ggplot(aes(x = dow)) +
  stat_summary(aes(y = daily_gross_bookings, colour = 'gross booking'), fun.y = 'mean', geom = 'bar') +
  stat_summary(aes(y = daily_cxl_bookings, colour = 'cancelled booking'), fun.y = 'mean', geom = 'bar')
ncol = 2)
```







Cxl rate ~ Cum Gross Bookings (controlled for days prior)

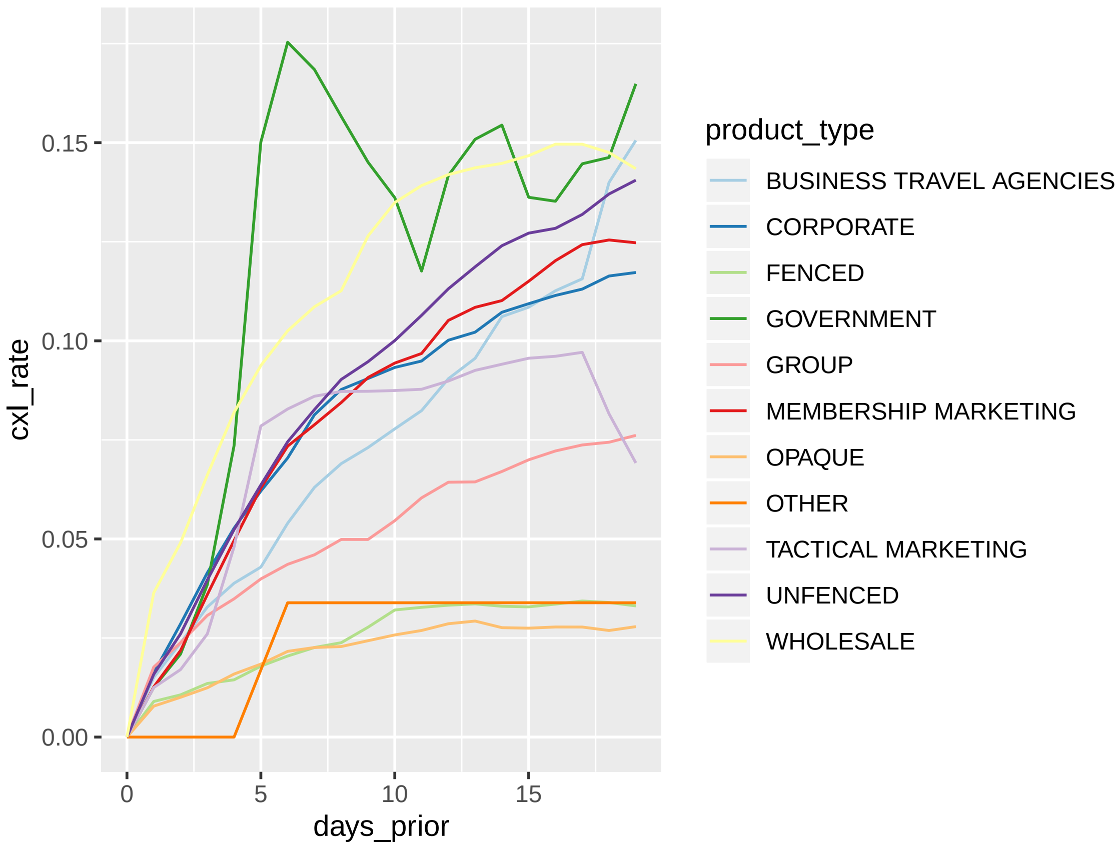
```
regr_cgb_cxl <- lm(cxl_rate ~ days_prior + cummulative_gross_bookings + product_type_regroup, data = tra
summary(regr_cgb_cxl)
```

```
##
## Call:
## lm(formula = cxl_rate ~ days_prior + cummulative_gross_bookings +
##     product_type_regroup, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20364 -0.07611 -0.02746  0.02558  1.00268
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.162e-01  1.871e-03   62.12   <2e-16 ***
## days_prior                      1.457e-03  4.362e-05   33.40   <2e-16 ***
## cummulative_gross_bookings     -1.123e-04  8.522e-06  -13.18   <2e-16 ***
## product_type_regroupMed_cxl    -5.871e-02  1.774e-03  -33.10   <2e-16 ***
## product_type_regroupLow_cxl    -1.261e-01  1.776e-03  -70.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.1544 on 44647 degrees of freedom
## Multiple R-squared:  0.1329, Adjusted R-squared:  0.1328
## F-statistic:  1711 on 4 and 44647 DF,  p-value: < 2.2e-16
```

Method 2: Consider last 20 days prior - three groups

```r
train %>% filter(days_prior <20) %>%
  ggplot(aes(x = days_prior, color = product_type)) +
  stat_summary(aes(y = cxl_rate), fun.y = 'mean', geom = 'line')+
  scale_color_brewer(palette = "Paired")
```



Cancellation trend of each product type. This is the main criteria for regrouping Only focus on the last 20 days

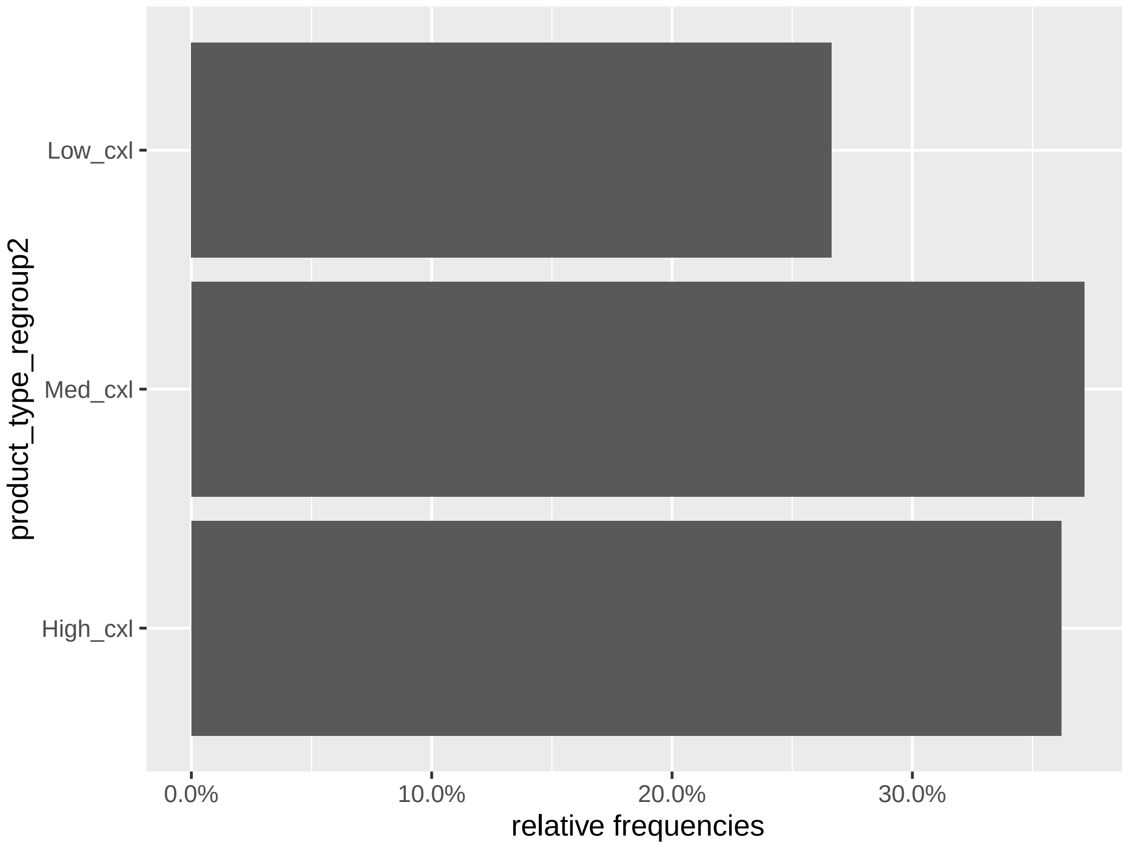
- High level of cancellation:
  - Business Travel Agencies (BTA)
  - Unfenced
  - Wholesale
  - Government
- Middle level of cancellation
  - Corporate
  - Membership Marketing
  - Tatical Marketing
  - Group
- Low lever of cancellation
  - Other

  - Opaque
  - Fenced

```
## Second Regrouping
# create product type group
high_cxl <- c('WHOLESALE', 'GOVERNMENT', 'BUSINESS TRAVEL AGENCIES', 'UNFENCED')
mid_cxl <- c('CORPORATE', 'MEMBERSHIP MARKETING','TACTICAL MARKETING', 'GROUP' )
low_cxl <- c('OPAQUE', 'OTHER','FENCED')

# Make Column in nyc dataset
nyc <- cbind(product_type_regroup2 = 'Other', nyc)
# Rename vars in product type level 2
nyc$product_type_regroup2 <- ifelse(nyc$product_type %in% high_cxl, 'High_cxl',
                             ifelse(nyc$product_type %in% mid_cxl, 'Med_cxl',
                                    ifelse(nyc$product_type %in% low_cxl, 'Low_cxl','Other')))

# Make Column in train dataset
train <- cbind(product_type_regroup2 = 'Other', train)
# Rename vars in product type level 2
train$product_type_regroup2 <- ifelse(train$product_type %in% high_cxl, 'High_cxl',
                             ifelse(train$product_type %in% mid_cxl, 'Med_cxl',
                                    ifelse(train$product_type %in% low_cxl, 'Low_cxl','Other')))

# Establish order
train$product_type_regroup2 <- factor(train$product_type_regroup2,
                  levels = c('High_cxl', 'Med_cxl', 'Low_cxl'))
nyc$product_type_regroup2 <- factor(nyc$product_type_regroup2,
                  levels = c('High_cxl', 'Med_cxl', 'Low_cxl'))
```

Check sample size of new grouping (2)

```
train %>%
ggplot(aes(x = product_type_regroup2)) +
  geom_bar(aes(y = (..count..)/sum(..count..)))+
  scale_y_continuous(labels=scales::percent) + coord_flip() +
  ylab("relative frequencies")
```



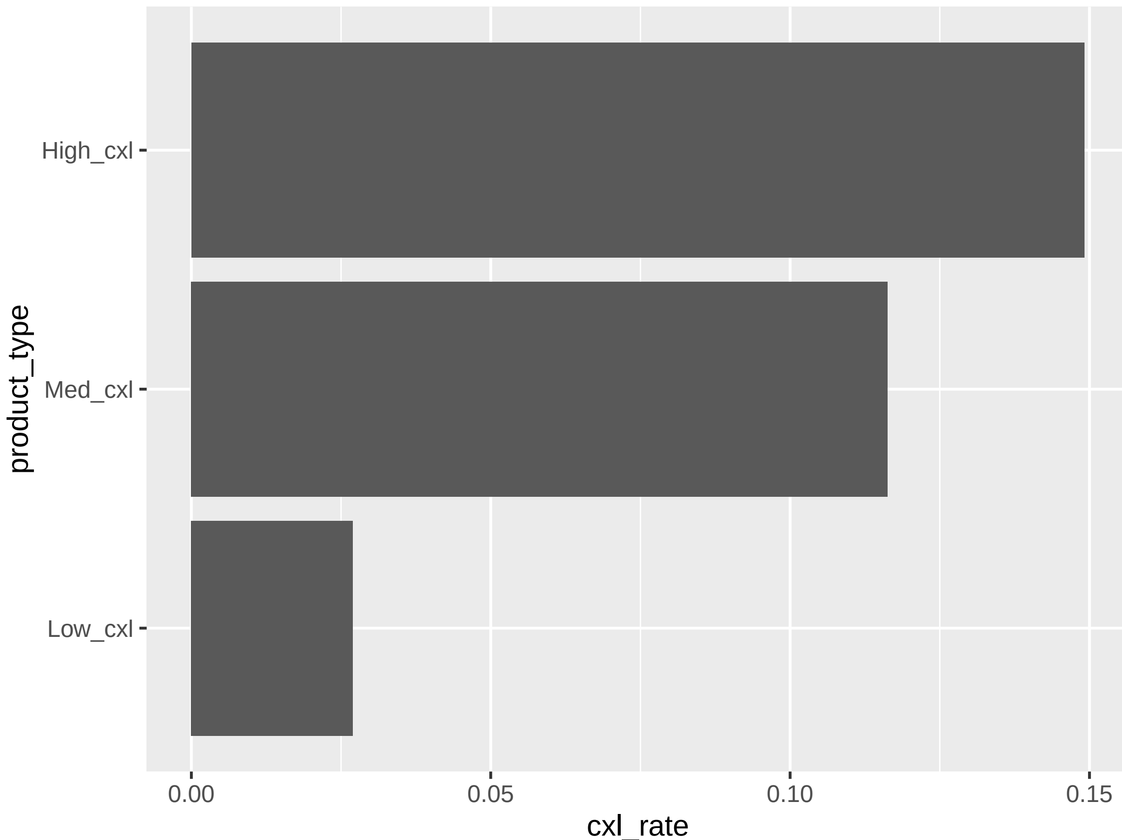
Check number of cancellations in each group (2)

```
kable(
train %>% group_by(product_type_regroup2) %>% summarise(cancellation = sum(daily_cxl_bookings)))  %>%
  kable_styling(bootstrap_options = "striped", full_width = F)
```

product_type_regroup2	cancellation
High_cxl	4667
Med_cxl	6927
Low_cxl	936

EDA with new grouping

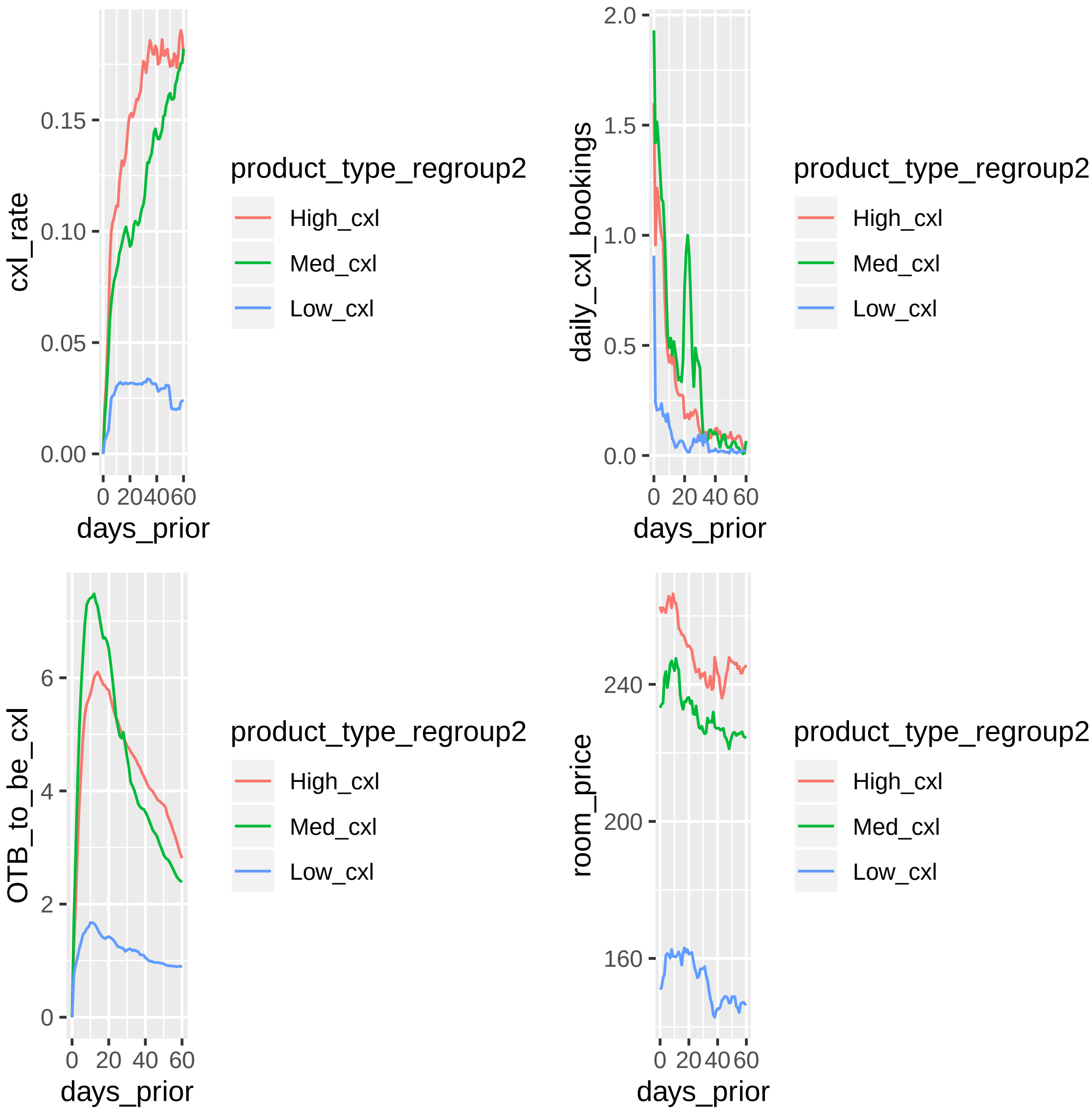
```
train %>% ggplot(aes(x = reorder(product_type_regroup2, cxl_rate), y = cxl_rate)) + stat_summary(fun.y =
```



Relationship with days prior

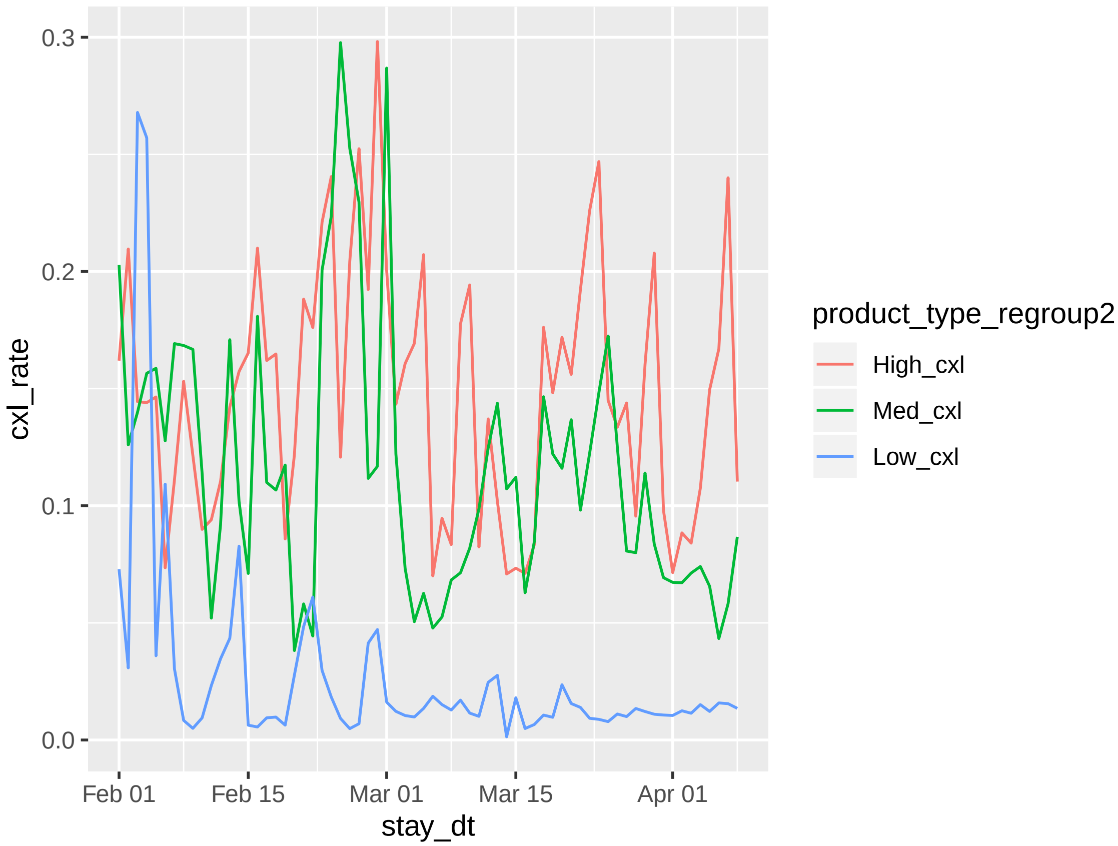
Cxl Rate does not differ as greatly as in Grouping 1

```
grid.arrange(
train %>%
  ggplot(aes(x = days_prior, color = product_type_regroup2)) +
  stat_summary(aes(y = cxl_rate), fun.y = 'mean', geom = 'line'),
train %>%
  ggplot(aes(x = days_prior, color = product_type_regroup2)) +
  stat_summary(aes(y = daily_cxl_bookings), fun.y = 'mean', geom = 'line'),
train %>%
  ggplot(aes(x = days_prior, color = product_type_regroup2)) +
  stat_summary(aes(y = OTB_to_be_cxl), fun.y = 'mean', geom = 'line'),
train %>%
  ggplot(aes(x = days_prior, color = product_type_regroup2)) +
  stat_summary(aes(y = room_price), fun.y = 'mean', geom = 'line'),
ncol = 2)
```



Relationship with stay_dt

```
train %>%
  ggplot(aes(x = stay_dt, color = product_type_regroup2)) +
  stat_summary(aes(y = cxl_rate), fun.y = 'mean', geom = 'line')
```

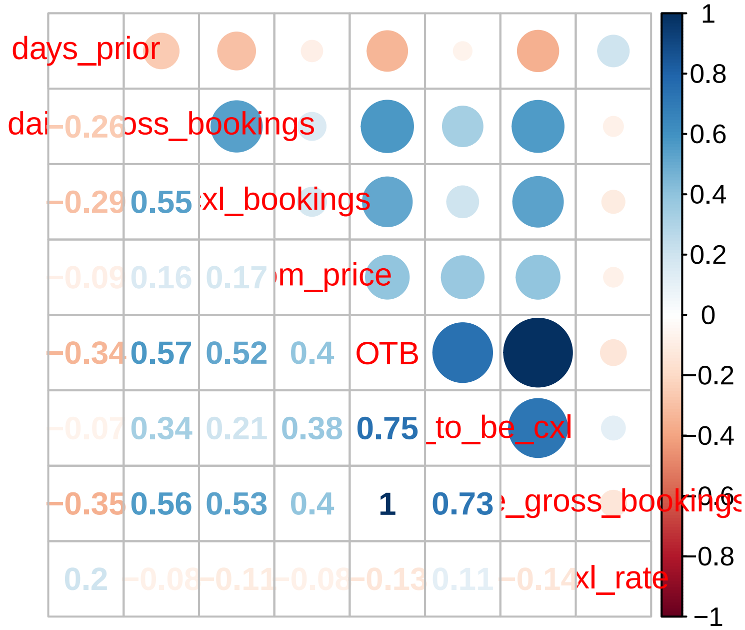


Correlational analysis

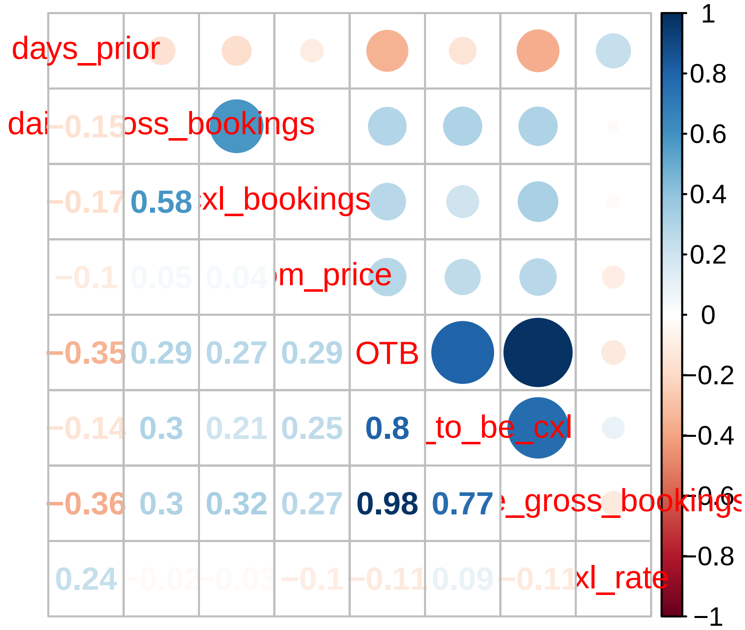
After regrouping, for different groups of product types, we can see the correlation change, which means that for different group, the impactor of cxl rate vary.

Room Price relationship to other dependent variables is more significant

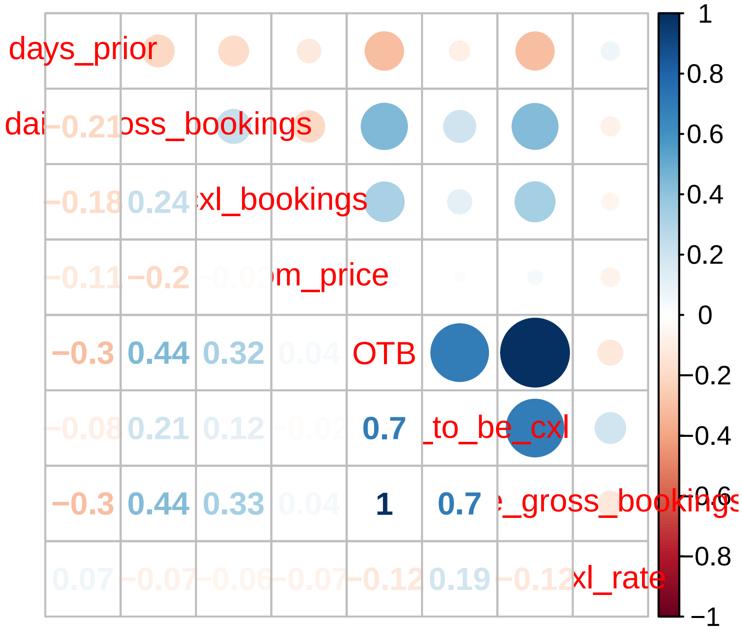
```
# Find correlation of quantitative variables

cor_plot <- train %>%
  filter(product_type_regroup2 == "High_cxl") %>%
  filter(room_price > 1, na.omit(room_price)) %>%  #Filter promotion room_price and missing value in ro
  select(days_prior, daily_gross_bookings,daily_cxl_bookings,room_price, OTB,  OTB_to_be_cxl, cummulati
a <- cor(cor_plot)

corrplot.mixed(a)
```




```
cor_plot <- train %>%
  filter(product_type_regroup2 == "Med_cxl") %>%
  filter(room_price > 1, na.omit(room_price)) %>%  #Filter promotion room_price and missing value in ro
  select(days_prior, daily_gross_bookings,daily_cxl_bookings,room_price, OTB,  OTB_to_be_cxl, cummulati
a <- cor(cor_plot)
corrplot.mixed(a)
```



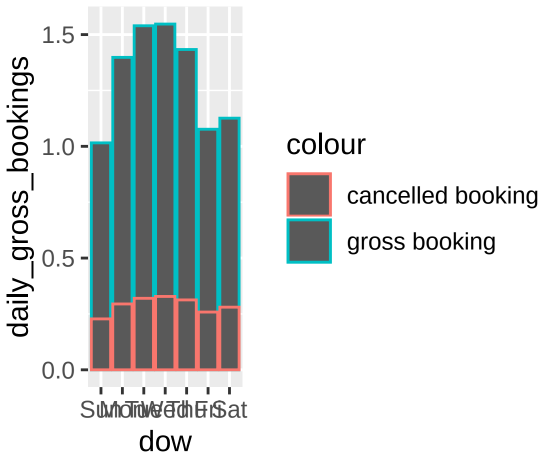

```
cor_plot <- train %>%
  filter(product_type_regroup2 == "Low_cxl") %>%
  filter(room_price > 1, na.omit(room_price)) %>%  #Filter promotion room_price and missing value in ro
  select(days_prior, daily_gross_bookings,daily_cxl_bookings,room_price, OTB,  OTB_to_be_cxl, cummulati
a <- cor(cor_plot)
corrplot.mixed(a)
```

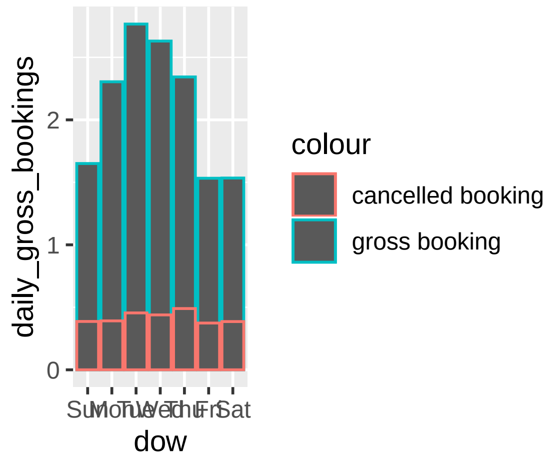


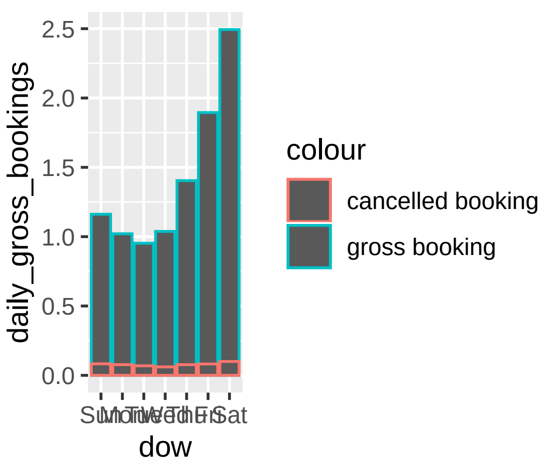
Relationship with DOW

We see that High and Medium Cxl group has similar pattern but Low Cxl group

```
grid.arrange(

train %>%
  filter(product_type_regroup2 == "High_cxl") %>%
  ggplot(aes(x = dow)) +
  stat_summary(aes(y = daily_gross_bookings, colour = 'gross booking'), fun.y = 'mean', geom = 'bar') +
  stat_summary(aes(y = daily_cxl_bookings, colour = 'cancelled booking'), fun.y = 'mean', geom = 'bar')
train %>%
  filter(product_type_regroup2== "Med_cxl") %>%
  ggplot(aes(x = dow)) +
  stat_summary(aes(y = daily_gross_bookings, colour = 'gross booking'), fun.y = 'mean', geom = 'bar') +
  stat_summary(aes(y = daily_cxl_bookings, colour = 'cancelled booking'), fun.y = 'mean', geom = 'bar')
train %>%
  filter(product_type_regroup2 == "Low_cxl") %>%
  ggplot(aes(x = dow)) +
  stat_summary(aes(y = daily_gross_bookings, colour = 'gross booking'), fun.y = 'mean', geom = 'bar') +
  stat_summary(aes(y = daily_cxl_bookings, colour = 'cancelled booking'), fun.y = 'mean', geom = 'bar')
ncol = 2)
```







Cxl rate ~ Cum Gross Bookings (controlled for days prior)

```r
regr_cgb_cxl2 <- lm(cxl_rate ~ days_prior + cummulative_gross_bookings + product_type_regroup2, data = t
summary(regr_cgb_cxl2)
```

```
##
## Call:
## lm(formula = cxl_rate ~ days_prior + cummulative_gross_bookings +
##     product_type_regroup2, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19793 -0.07536 -0.02285  0.02865  1.00588
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.075e-01  1.922e-03  55.907   <2e-16 ***
## days_prior                    1.508e-03  4.400e-05  34.266   <2e-16 ***
## cummulative_gross_bookings   -8.029e-05  8.603e-06  -9.332   <2e-16 ***
## product_type_regroup2Med_cxl -3.156e-02  1.726e-03 -18.283   <2e-16 ***
## product_type_regroup2Low_cxl -1.208e-01  1.887e-03 -64.036   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.1557 on 44647 degrees of freedom
## Multiple R-squared:  0.1183, Adjusted R-squared:  0.1182
## F-statistic:  1497 on 4 and 44647 DF,  p-value: < 2.2e-16
```

Method 3: Consider last 20 days prior - two groups only

Group 3 is based on Group2, but just simply combine high cxl group and Med cxl group into one group

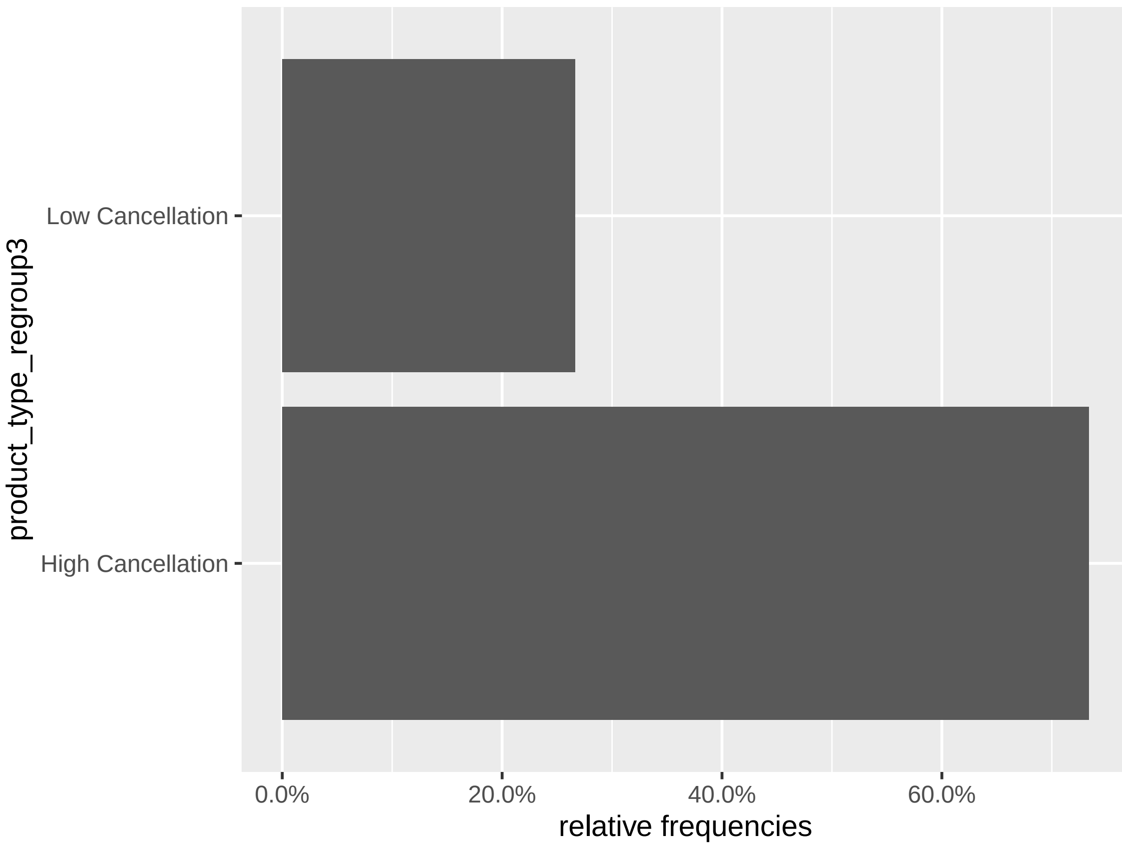
- High level of cancellation:
- Business Travel Agencies (BTA)
- Unfenced
- Whlesale
- Government
- Corporate
- Membership Marketing
- Tatical Marketing
- Group
- Low lever of cancellation
- Other
- Opaque
- Fenced

```
## Third Regrouping
# create product type group
high_cxl <- c('WHOLESALE', 'GOVERNMENT', 'BUSINESS TRAVEL AGENCIES', 'UNFENCED','CORPORATE', 'MEMBERSHI
low_cxl <- c('OPAQUE', 'OTHER','FENCED')

# Make Column in nyc dataset
nyc <- cbind(product_type_regroup3 = 'Other', nyc)
# Rename vars in product type level 3
nyc$product_type_regroup3 <- ifelse(nyc$product_type %in% high_cxl, 'High Cancellation',  "Low Cancella

# Make Column in train dataset
train <- cbind(product_type_regroup3 = 'Other', train)
# Rename vars in product type level 3
train$product_type_regroup3 <- ifelse(train$product_type %in% high_cxl, 'High Cancellation',"Low Cancel


# Establish order
train$product_type_regroup3 <- factor(train$product_type_regroup3,
                  levels = c('High Cancellation',  'Low Cancellation'))
nyc$product_type_regroup3 <- factor(nyc$product_type_regroup3,
                  levels = c('High Cancellation',  'Low Cancellation'))
```

Check sample size

```
train %>%
ggplot(aes(x = product_type_regroup3)) +
  geom_bar(aes(y = (..count..)/sum(..count..)))+
```

```r
  scale_y_continuous(labels=scales::percent) + coord_flip() +
  ylab("relative frequencies")
```



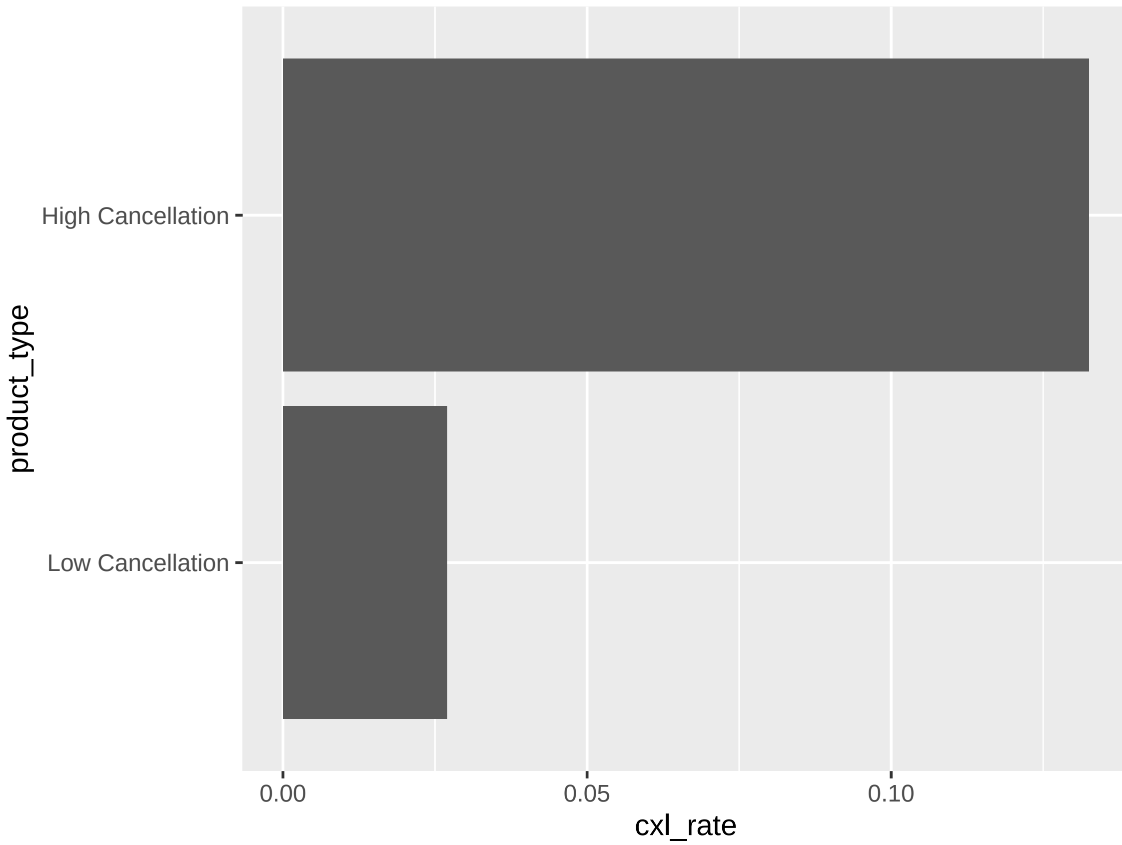
Check number of cancellations in each group

```r
kable(
train %>% group_by(product_type_regroup3) %>% summarise(cancellation = sum(daily_cxl_bookings)),
format = 'html')
```

product_type_regroup3	cancellation
High Cancellation	11594
Low Cancellation	936

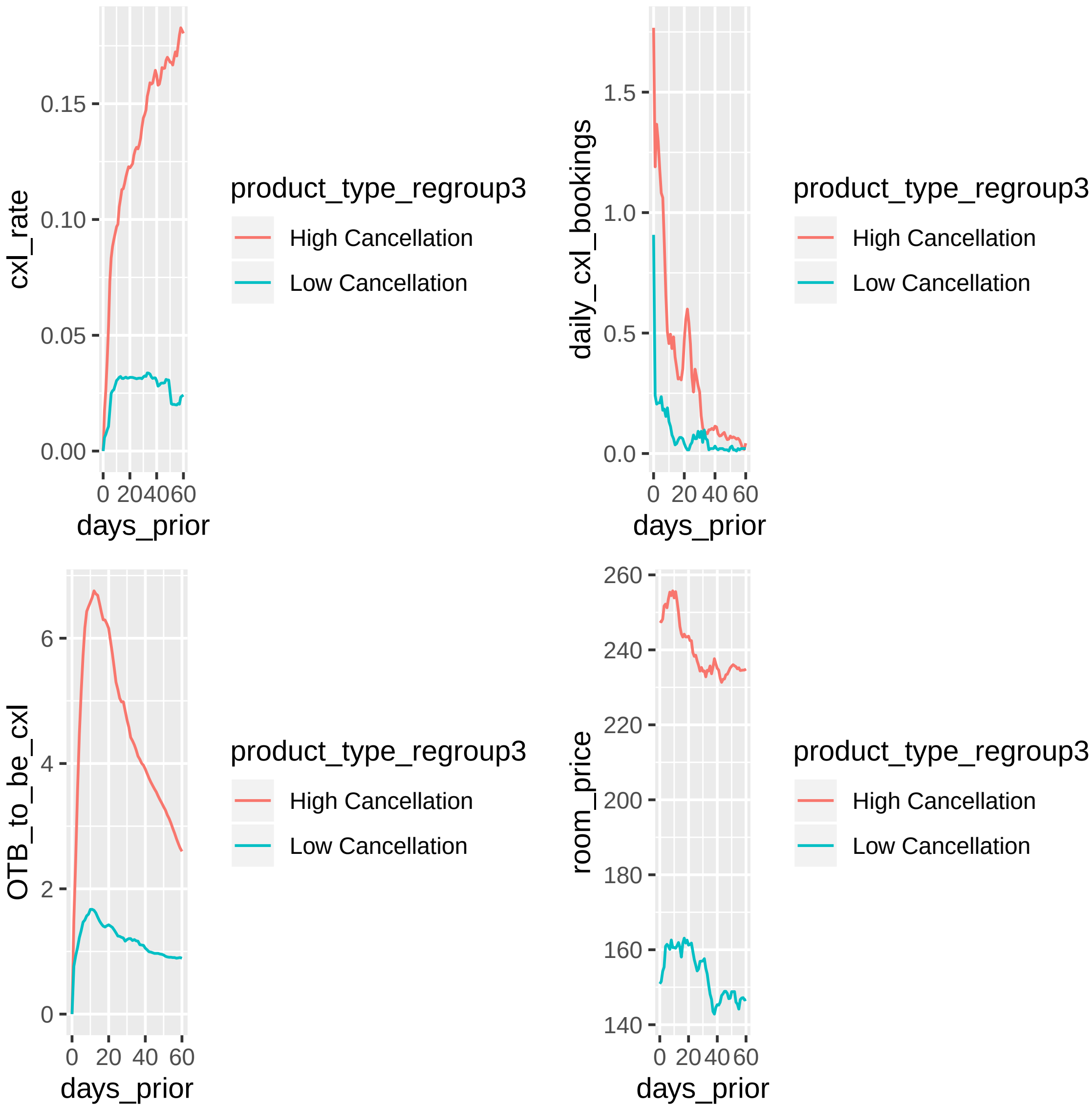
EDA with new grouping

```r
train %>% ggplot(aes(x = reorder(product_type_regroup3, cxl_rate), y = cxl_rate)) + stat_summary(fun.y =
```



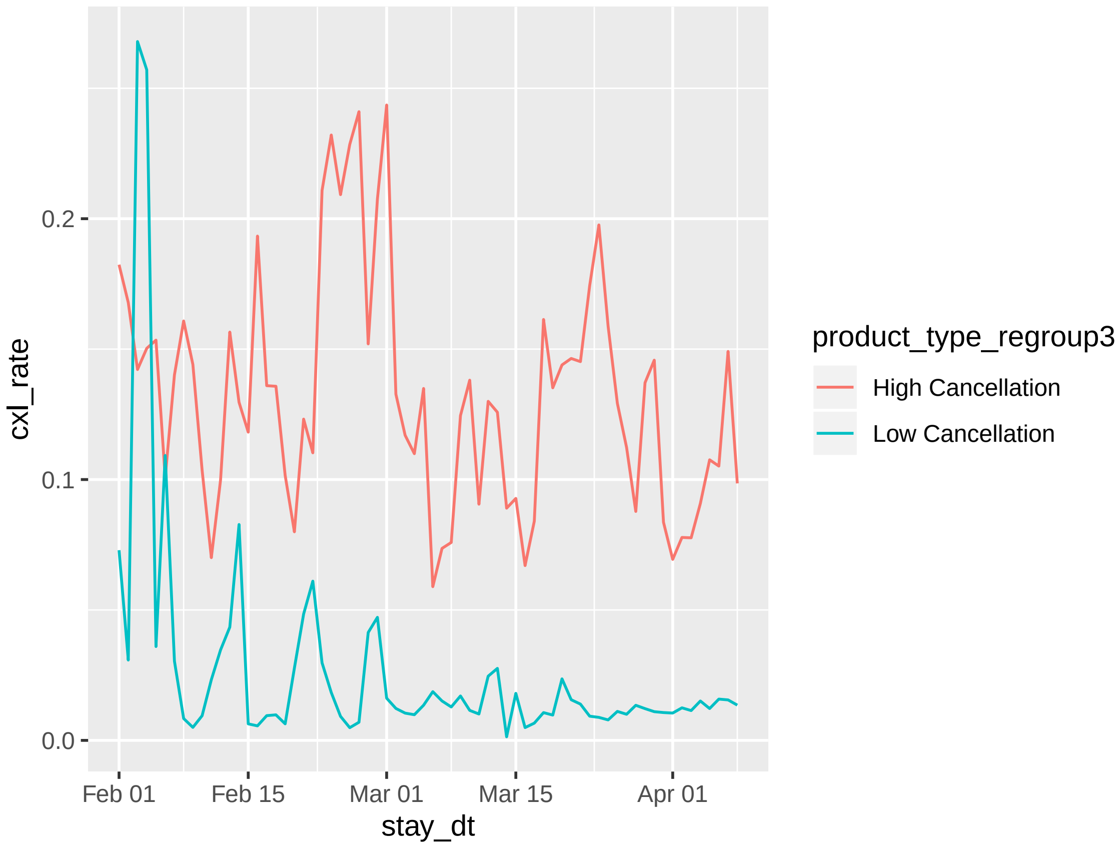
Relationship with days prior

```
grid.arrange(
train %>%
  ggplot(aes(x = days_prior, color = product_type_regroup3)) +
  stat_summary(aes(y = cxl_rate), fun.y = 'mean', geom = 'line'),
train %>%
  ggplot(aes(x = days_prior, color = product_type_regroup3)) +
  stat_summary(aes(y = daily_cxl_bookings), fun.y = 'mean', geom = 'line'),
train %>%
  ggplot(aes(x = days_prior, color = product_type_regroup3)) +
  stat_summary(aes(y = OTB_to_be_cxl), fun.y = 'mean', geom = 'line'),
train %>%
  ggplot(aes(x = days_prior, color = product_type_regroup3)) +
  stat_summary(aes(y = room_price), fun.y = 'mean', geom = 'line'),
ncol = 2)
```



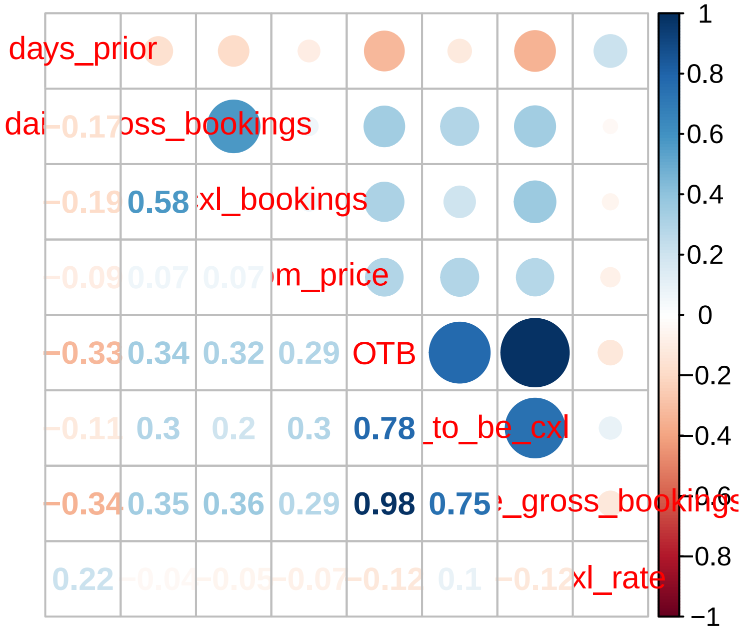
Relationship with stay_dt

```
train %>%
  ggplot(aes(x = stay_dt, color = product_type_regroup3)) +
  stat_summary(aes(y = cxl_rate), fun.y = 'mean', geom = 'line')
```

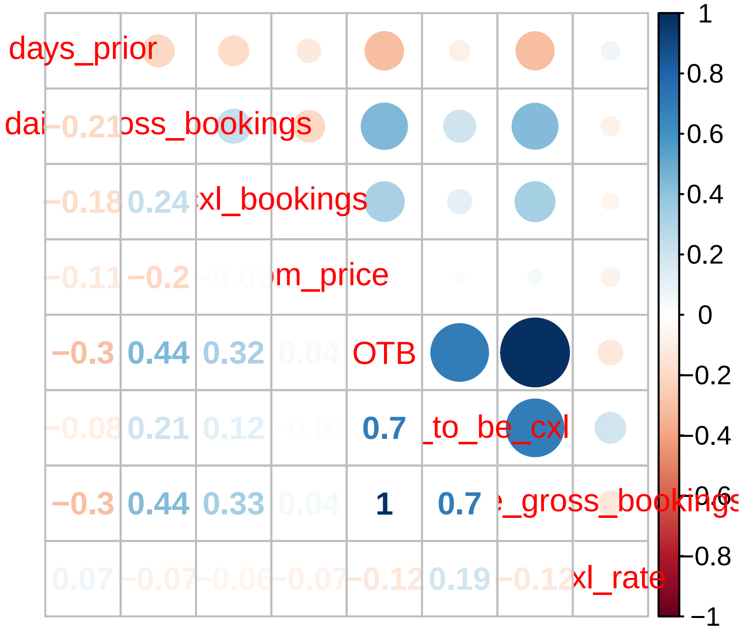


Correlational analysis

```
# Find correlation of quantitative variables

cor_plot <- train %>%
  filter(product_type_regroup3 == "High Cancellation") %>%
  filter(room_price > 1, na.omit(room_price)) %>%  #Filter promotion room_price and missing value in ro
  select(days_prior, daily_gross_bookings,daily_cxl_bookings,room_price, OTB,  OTB_to_be_cxl,  cummulati
a <- cor(cor_plot)
corrplot.mixed(a)
```



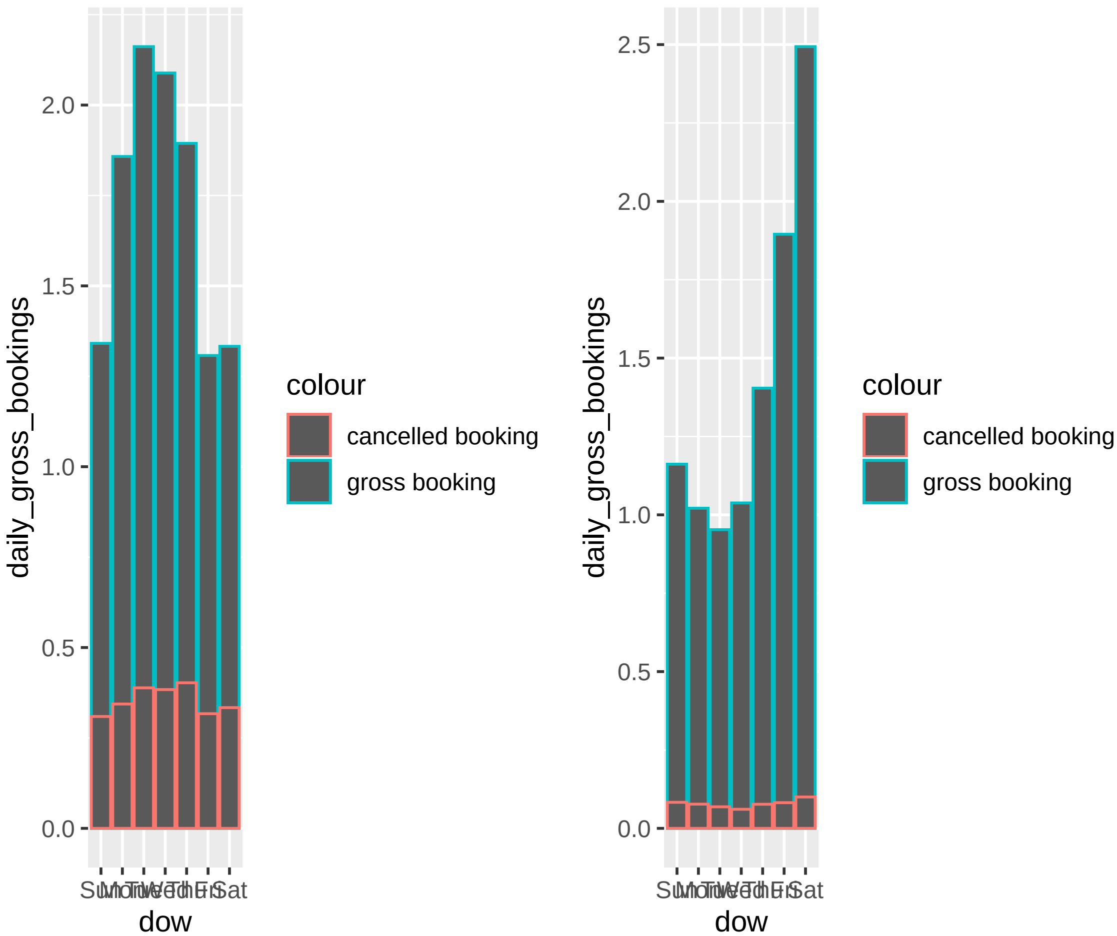
```
cor_plot <- train %>%
  filter(product_type_regroup3 == "Low Cancellation") %>%
  filter(room_price > 1, na.omit(room_price)) %>%  #Filter promotion room_price and missing value in ro
  select(days_prior, daily_gross_bookings,daily_cxl_bookings,room_price, OTB,  OTB_to_be_cxl, cummulati
a <- cor(cor_plot)
corrplot.mixed(a)
```



After regrouping, for different groups of product types, we can see the correlation change, which means that for different group, the impactor of cxl rate vary.

Relationship with DOW

```r
grid.arrange(

train %>%
  filter(product_type_regroup3 == "High Cancellation") %>%
  ggplot(aes(x = dow)) +
  stat_summary(aes(y = daily_gross_bookings, colour = 'gross booking'), fun.y = 'mean', geom = 'bar') +
  stat_summary(aes(y = daily_cxl_bookings, colour = 'cancelled booking'), fun.y = 'mean', geom = 'bar')

train %>%
  filter(product_type_regroup3 == "Low Cancellation") %>%
  ggplot(aes(x = dow)) +
  stat_summary(aes(y = daily_gross_bookings, colour = 'gross booking'), fun.y = 'mean', geom = 'bar') +
  stat_summary(aes(y = daily_cxl_bookings, colour = 'cancelled booking'), fun.y = 'mean', geom = 'bar')
ncol = 2)
```



Cxl rate ~ Cum Gross Bookings (controlled for days prior)

```r
regr_cgb_cxl3 <- lm(cxl_rate ~ days_prior + cummulative_gross_bookings + product_type_regroup3, data =
summary(regr_cgb_cxl3)
```

```
##
## Call:
## lm(formula = cxl_rate ~ days_prior + cummulative_gross_bookings +
##     product_type_regroup3, data = train)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18201 -0.07581 -0.02267  0.02975  1.00459
##
## Coefficients:
##                                       Estimate Std. Error t value
## (Intercept)                          9.277e-02  1.753e-03   52.93
## days_prior                           1.487e-03  4.415e-05   33.69
## cummulative_gross_bookings          -9.316e-05  8.606e-06  -10.82
## product_type_regroup3Low Cancellation -1.047e-01  1.675e-03  -62.53
##                                     Pr(>|t|)
## (Intercept)                           <2e-16 ***
## days_prior                            <2e-16 ***
## cummulative_gross_bookings            <2e-16 ***
## product_type_regroup3Low Cancellation   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1563 on 44648 degrees of freedom
## Multiple R-squared:  0.1117, Adjusted R-squared:  0.1116
## F-statistic:  1871 on 3 and 44648 DF,  p-value: < 2.2e-16
```