Hotel Cancellation Study

Leo Shi

2024-05-14

```r
#packages
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
## Warning: package 'tibble' was built under R version 4.1.3
```

```
## Warning: package 'tidyr' was built under R version 4.1.3
```

```
## Warning: package 'purrr' was built under R version 4.1.3
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
## Warning: package 'stringr' was built under R version 4.1.3
```

```
## Warning: package 'forcats' was built under R version 4.1.3
```

```
## Warning: package 'lubridate' was built under R version 4.1.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v purrr     1.0.1
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```r
library(ggplot2)
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.1.3
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(dplyr)
library(car)
```

```
## Warning: package 'car' was built under R version 4.1.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.1.3
```

```
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```r
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.1.3
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
##
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(arm)
```

```
## Loading required package: Matrix
```

```
## Warning: package 'Matrix' was built under R version 4.1.3
```

```
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
## Loading required package: lme4
```

```
## Warning: package 'lme4' was built under R version 4.1.3
```

```
##
## arm (Version 1.14-4, built: 2024-4-1)
##
## Working directory is C:/Users/Leo Shi/Desktop/Homework Spring 2024
##
##
## Attaching package: 'arm'
##
## The following object is masked from 'package:car':
##
##     logit
```

```
library(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 4.1.3
```

```
library(ggfortify)
```

```
#Display the initial dataset
Hotel_bookings <- read_csv("Hotel_bookings.csv")
```

```
## Rows: 119390 Columns: 32
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (14): hotel, arrival_date_month, meal, country, market_segment, distribu...
## dbl (18): is_canceled, lead_time, arrival_date_year, arrival_date_week_numbe...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(Hotel_bookings)
```

```
## # A tibble: 6 x 32
##   hotel        is_canceled lead_time arrival_date_year arrival_date_month
##   <chr>              <dbl>     <dbl>             <dbl> <chr>
## 1 Resort Hotel           0       342              2015 July
## 2 Resort Hotel           0       737              2015 July
## 3 Resort Hotel           0         7              2015 July
## 4 Resort Hotel           0        13              2015 July
## 5 Resort Hotel           0        14              2015 July
## 6 Resort Hotel           0        14              2015 July
## # i 27 more variables: arrival_date_week_number <dbl>,
## #   arrival_date_day_of_month <dbl>, stays_in_weekend_nights <dbl>,
## #   stays_in_week_nights <dbl>, adults <dbl>, children <dbl>, babies <dbl>,
## #   meal <chr>, country <chr>, market_segment <chr>,
## #   distribution_channel <chr>, is_repeated_guest <dbl>,
## #   previous_cancellations <dbl>, previous_bookings_not_canceled <dbl>,
## #   reserved_room_type <chr>, assigned_room_type <chr>, ...
```

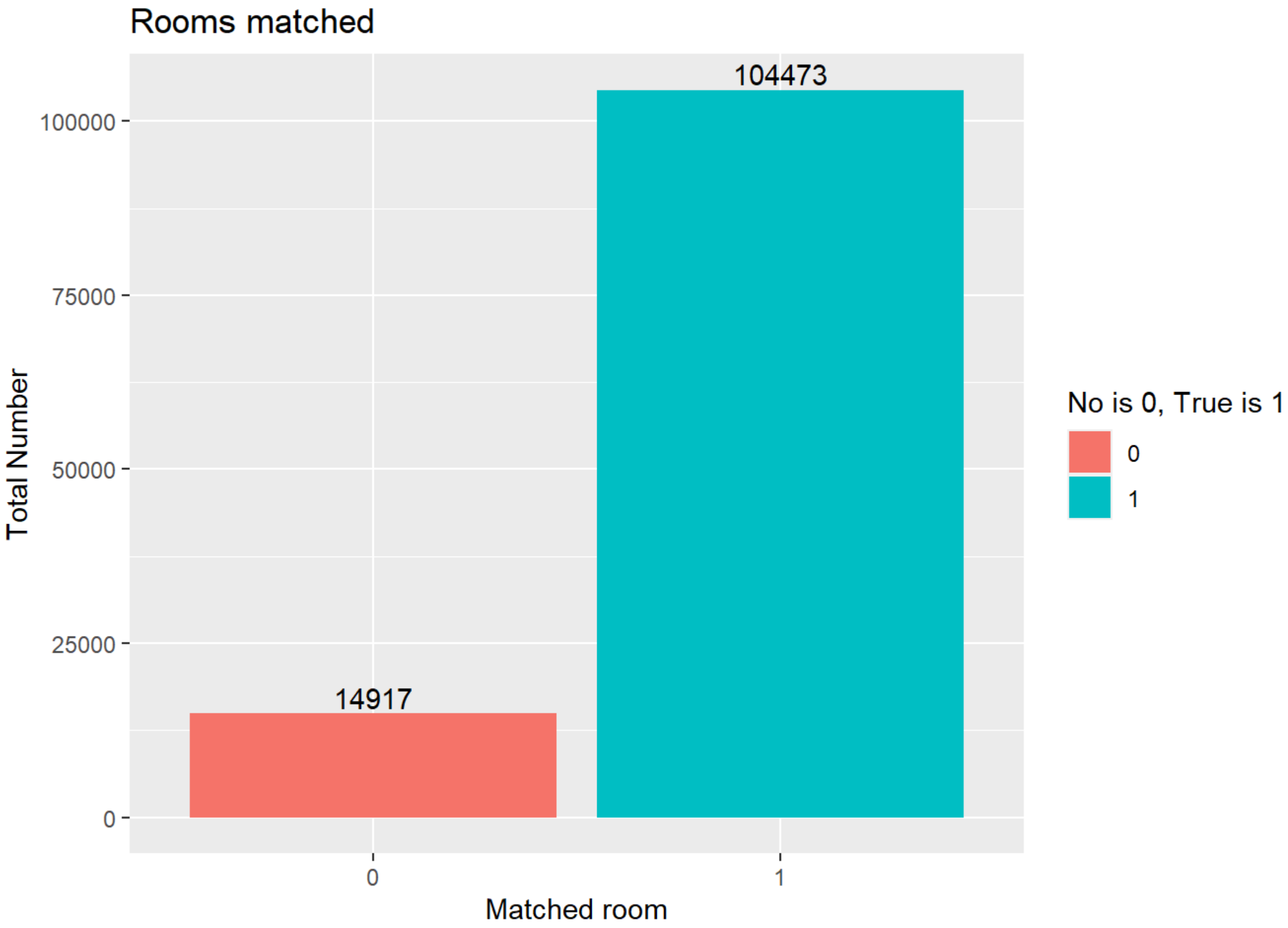
```
#EDA exploration for categorical
#Matching rooms
Hotel_bookings$matched <- ifelse(Hotel_bookings$reserved_room_type == Hotel_bookings$assigned_ro
om_type, "1", "0")
head(as.character(Hotel_bookings$matched))
```

```
## [1] "1" "1" "0" "1" "1" "1"
```

```
matched_room <- table(Hotel_bookings$matched)
matched_room_df <- as.data.frame(matched_room)
colnames(matched_room_df) <- c('Matched room', 'Total Number')
matched_room_df
```

```
##   Matched room Total Number
## 1            0        14917
## 2            1       104473
```

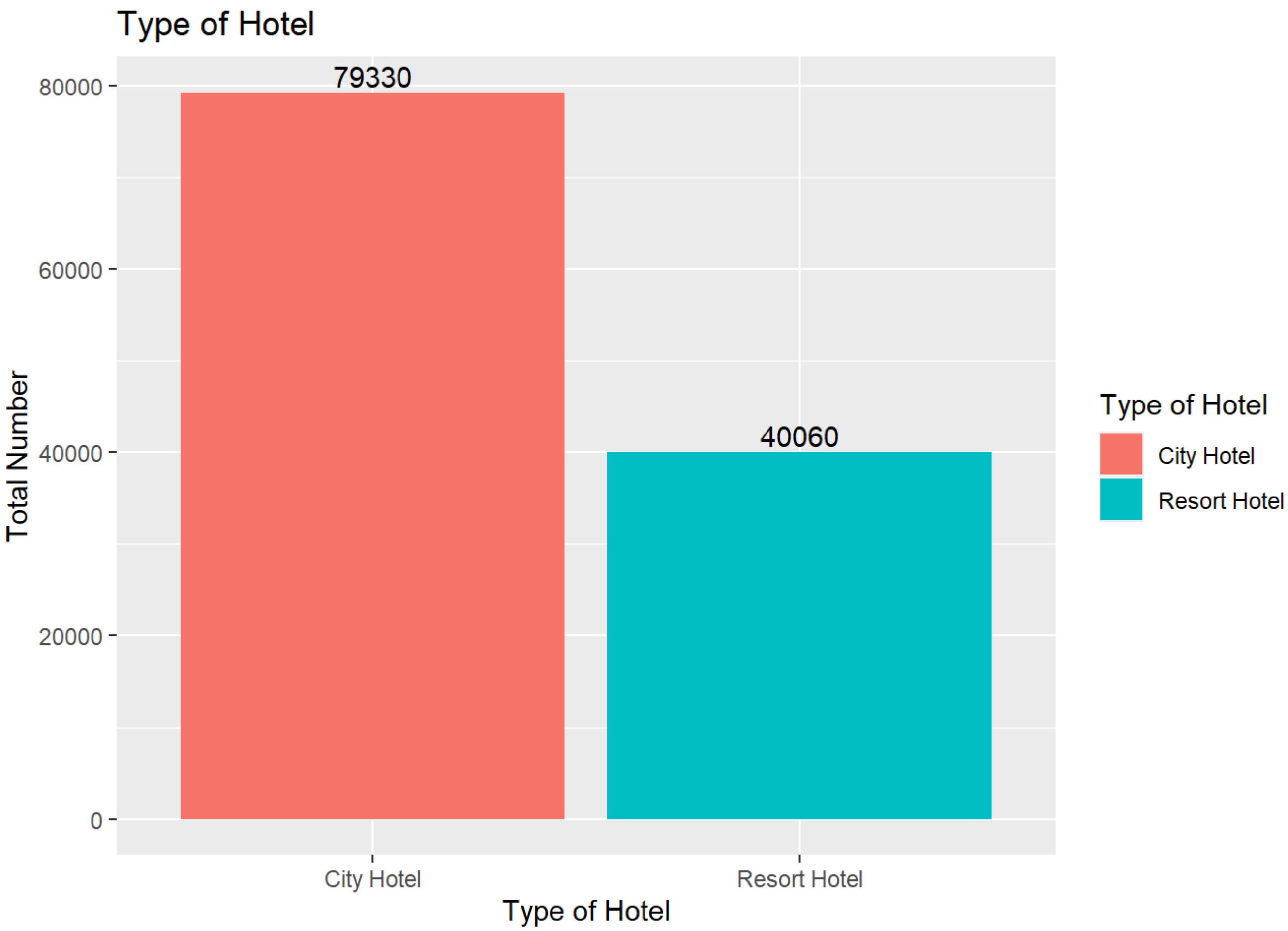
```
# Create the pie chart for market segment
matched_room_chart<- ggplot(data=matched_room_df, aes(x=`Matched room`, y=`Total Number`, fill=`
Matched room`))+
  geom_bar(stat="identity")+
    labs(title = "Rooms matched", fill = "No is 0, True is 1")+
  geom_text(aes(label=`Total Number`), position=position_dodge(width=0.9), vjust=-0.25)
matched_room_chart
```



```
hotel_types <- table(Hotel_bookings$hotel)
hotel_types_df <- as.data.frame(hotel_types)
colnames(hotel_types_df) <- c('Type of Hotel', 'Total Number')
hotel_types_df
```

```
##   Type of Hotel Total Number
## 1    City Hotel        79330
## 2  Resort Hotel        40060
```

```
# Create the pie chart for market segment
hotel_types_chart<- ggplot(data=hotel_types_df, aes(x=`Type of Hotel`, y=`Total Number`, fill=`T
ype of Hotel`))+
  geom_bar(stat="identity")+
    labs(title = "Type of Hotel")+
  geom_text(aes(label=`Total Number`), position=position_dodge(width=0.9), vjust=-0.25)
hotel_types_chart
```

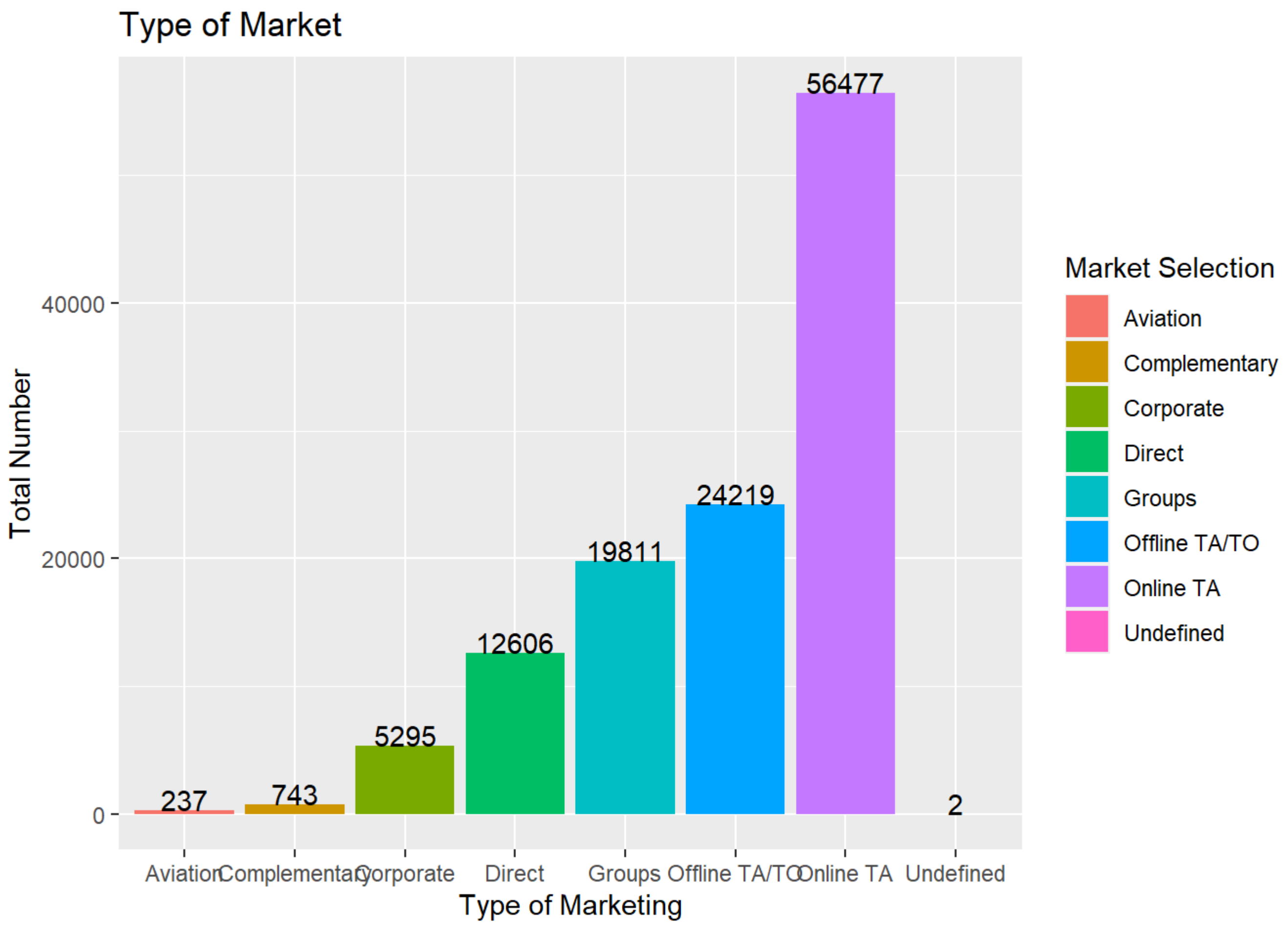


```
hotel_market <- table(Hotel_bookings$market_segment)
hotel_market_df <- as.data.frame(hotel_market)
colnames(hotel_market_df) <- c('Type of Marketing', 'Total Number')
hotel_market_df
```

```
##   Type of Marketing Total Number
## 1          Aviation          237
## 2     Complementary          743
## 3         Corporate         5295
## 4            Direct        12606
## 5            Groups        19811
## 6     Offline TA/TO        24219
## 7         Online TA        56477
## 8         Undefined            2
```

```
# Create the Bar chart for market segment

hotel_market_chart <- ggplot(data=hotel_market_df, aes(x=`Type of Marketing`, y=`Total Number`,
fill=`Type of Marketing`))+
  geom_bar(stat="identity")+
    labs(title = "Type of Market", fill = "Market Selection")+
  geom_text(aes(label=`Total Number`), position=position_dodge(width=1.2), vjust=0)
hotel_market_chart
```

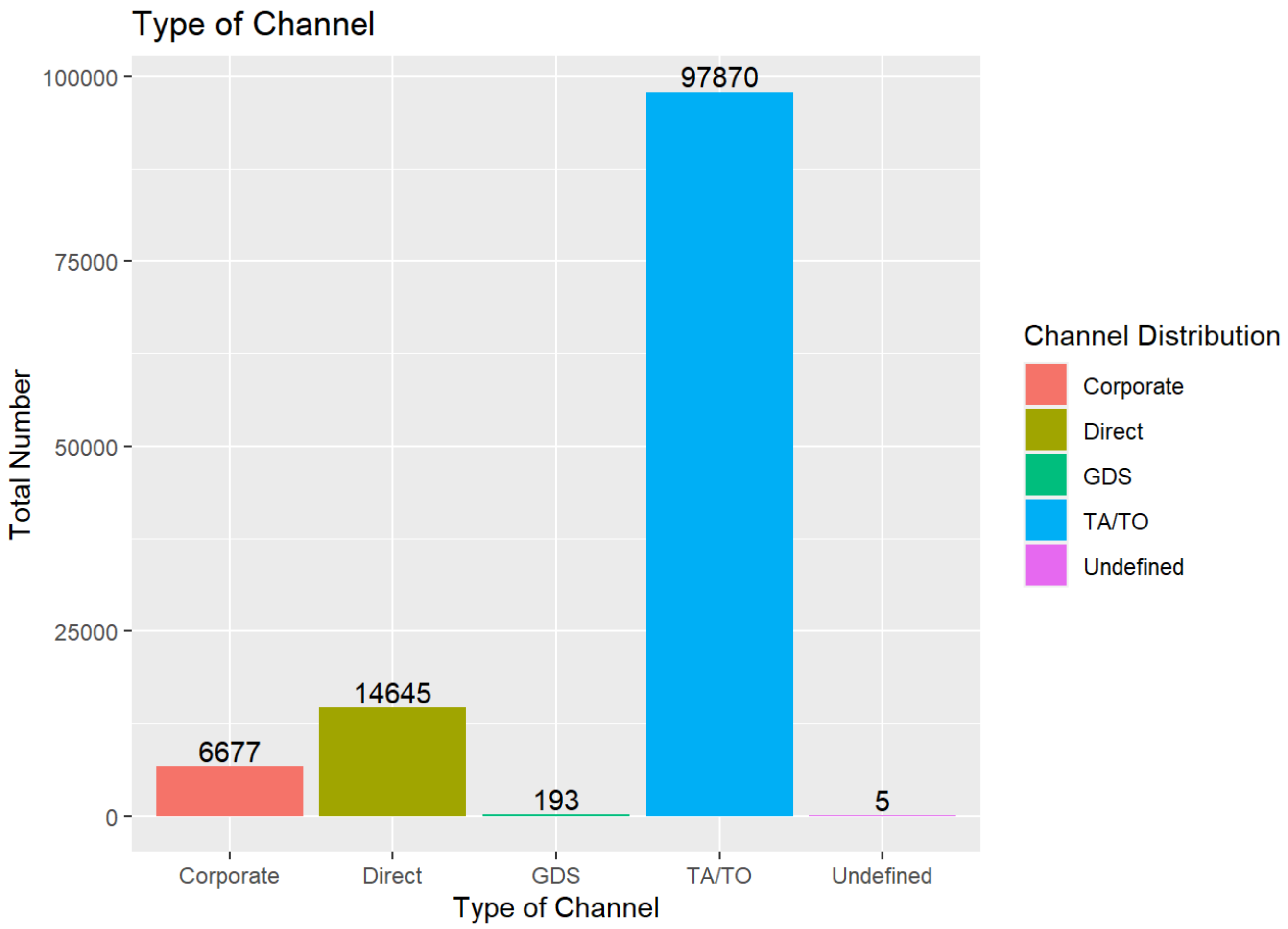
```
## Warning: `position_dodge()` requires non-overlapping x intervals
```



```
#Distribution channel
hotel_channel <- table(Hotel_bookings$distribution_channel)
hotel_channel_df <- as.data.frame(hotel_channel)
colnames(hotel_channel_df) <- c('Type of Channel', 'Total Number')
hotel_channel_df
```

```
##   Type of Channel Total Number
## 1       Corporate         6677
## 2          Direct        14645
## 3             GDS          193
## 4           TA/TO        97870
## 5       Undefined            5
```

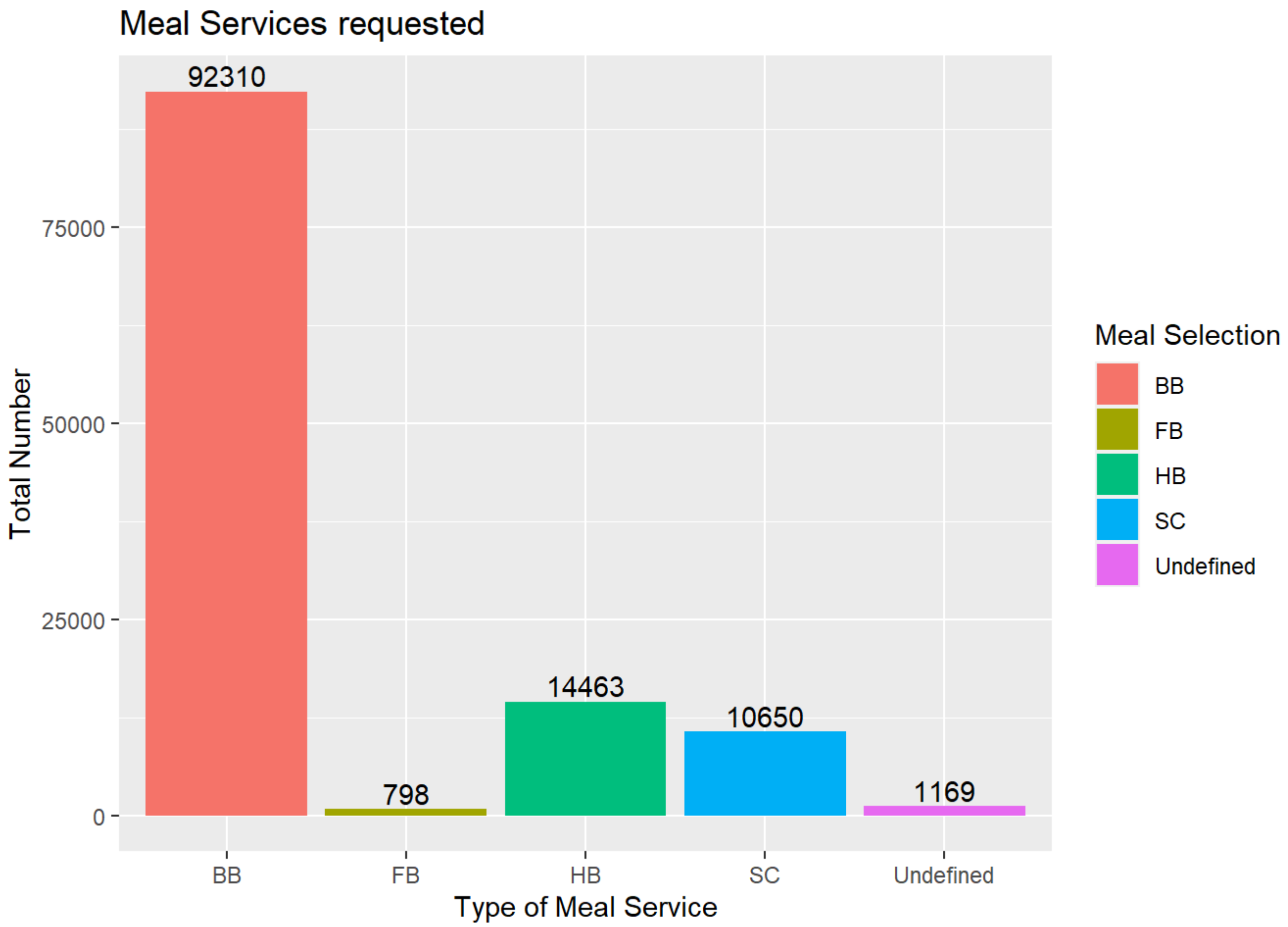
```
# Create the pie chart for channel segment
hotel_channel_chart <- ggplot(data=hotel_channel_df, aes(x=`Type of Channel`, y=`Total Number`,
fill=`Type of Channel`))+
  geom_bar(stat="identity")+
    labs(title = "Type of Channel", fill = "Channel Distribution")+
  geom_text(aes(label=`Total Number`), position=position_dodge(width=0.9), vjust=-0.25)
hotel_channel_chart
```



```
#type of meals
Meals <- table(Hotel_bookings$meal)
Meals_df <- as.data.frame(Meals)
colnames(Meals_df) <- c('Type of Meal Service', 'Total Number')
Meals_df
```

```
##   Type of Meal Service Total Number
## 1                   BB        92310
## 2                   FB          798
## 3                   HB        14463
## 4                   SC        10650
## 5            Undefined         1169
```

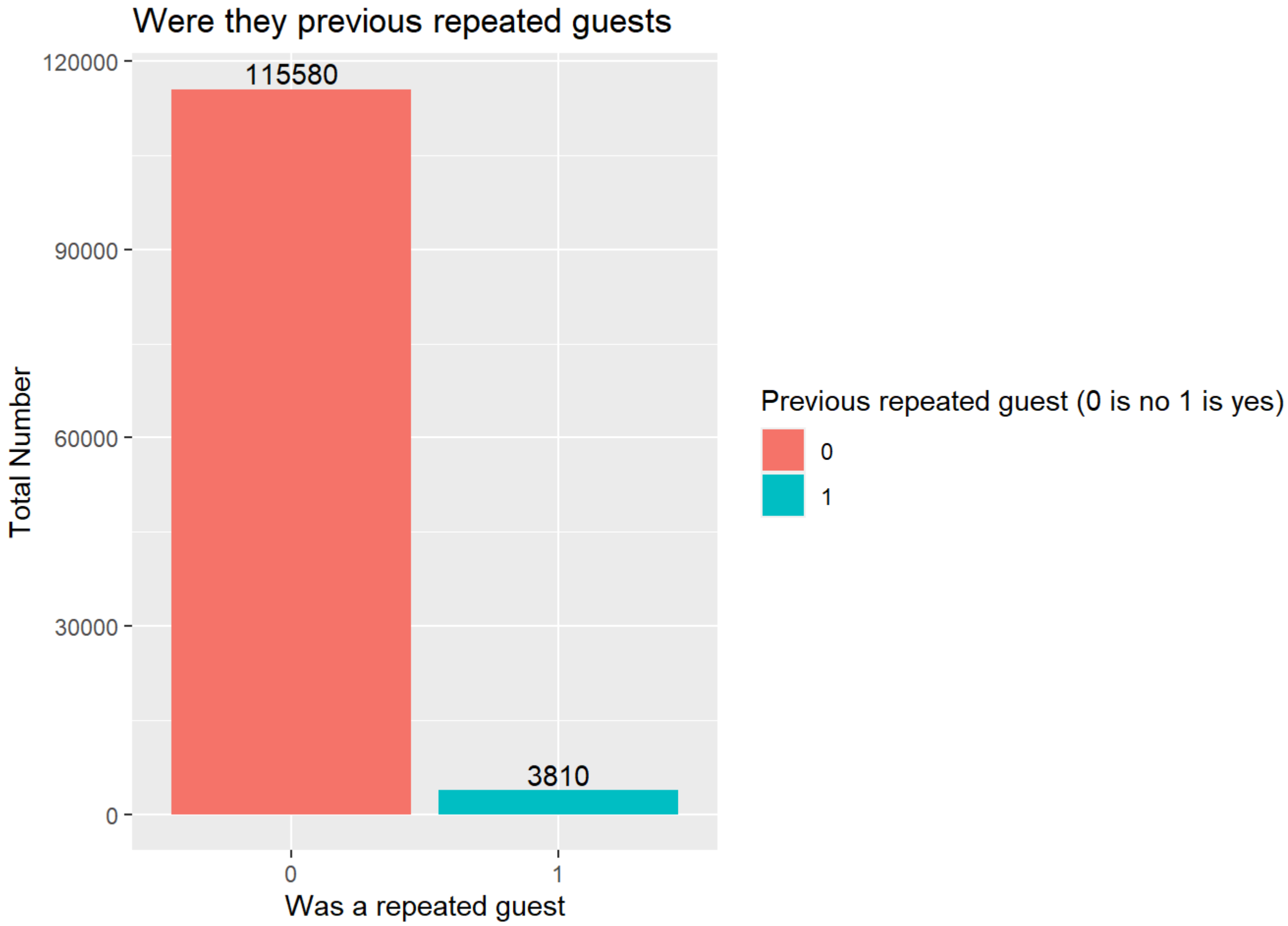
```
# Create the pie chart for Meals

Meals_chart <-ggplot(data=Meals_df, aes(x=`Type of Meal Service`, y=`Total Number`, fill=`Type o
f Meal Service`))+
  geom_bar(stat="identity")+
    labs(title = "Meal Services requested", fill = "Meal Selection")+
  geom_text(aes(label=`Total Number`), position=position_dodge(width=0.9), vjust=-0.25)
Meals_chart
```



```
#was a repeated guest
repeated_guest <- table(Hotel_bookings$is_repeated_guest)
repeated_guest_df <- as.data.frame(repeated_guest)
colnames(repeated_guest_df) <- c('Was a repeated guest', 'Total Number')
repeated_guest_df
```

```
##   Was a repeated guest Total Number
## 1                    0       115580
## 2                    1         3810
```

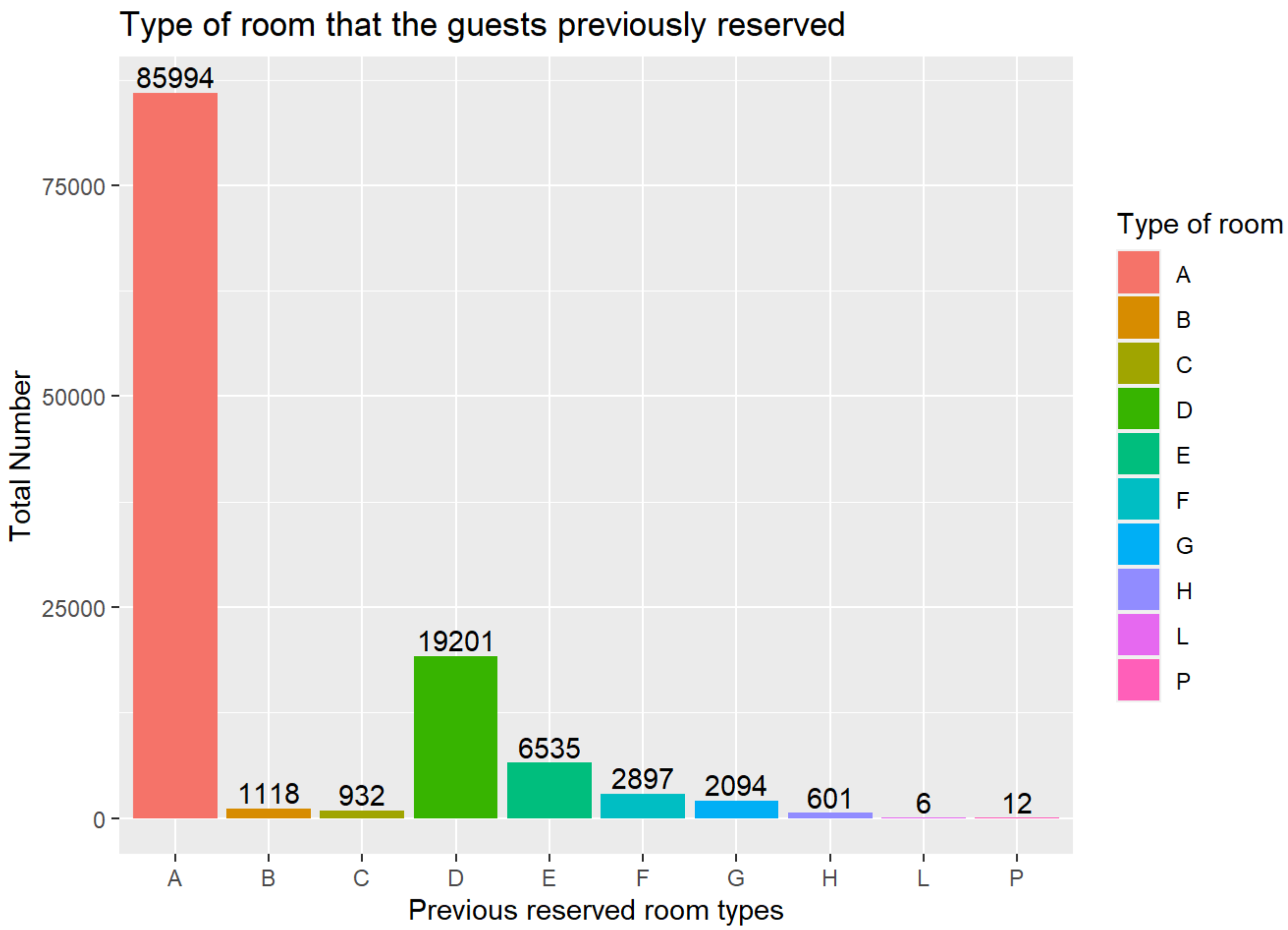
```
# Create the pie chart for market segment

repeated_guest_chart <- ggplot(data=repeated_guest_df, aes(x=`Was a repeated guest`, y=`Total Number`, fill=`Was a repeated guest`))+
  geom_bar(stat="identity")+
    labs(title = "Were they previous repeated guests", fill = "Previous repeated guest (0 is no 1 is yes)")+
  geom_text(aes(label=`Total Number`), position=position_dodge(width=0.9), vjust=-0.25)
repeated_guest_chart
```



```
#reserved_room type
reserved_room_type <- table(Hotel_bookings$reserved_room_type)
reserved_room_type_df <- as.data.frame(reserved_room_type)
colnames(reserved_room_type_df) <- c('Previous reserved room types', 'Total Number')
reserved_room_type_df
```

```
##    Previous reserved room types Total Number
## 1                             A        85994
## 2                             B         1118
## 3                             C          932
## 4                             D        19201
## 5                             E         6535
## 6                             F         2897
## 7                             G         2094
## 8                             H          601
## 9                             L            6
## 10                            P           12
```

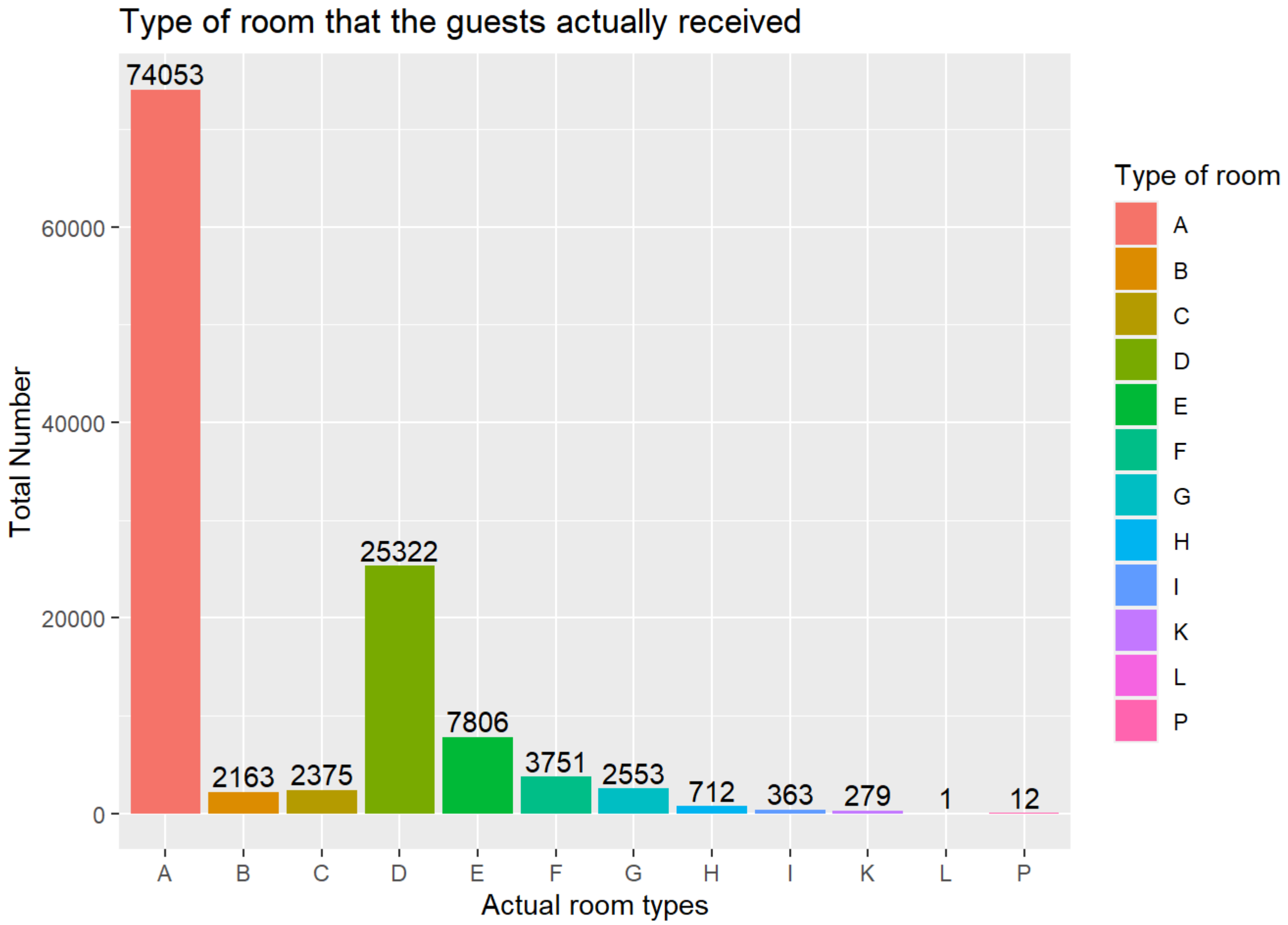
```
# Create the pie chart for market segment

reserved_room_chart <-ggplot(data=reserved_room_type_df, aes(x=`Previous reserved room types`, y
=`Total Number`, fill=`Previous reserved room types`))+
  geom_bar(stat="identity")+
    labs(title = "Type of room that the guests previously reserved", fill = "Type of room")+
  geom_text(aes(label=`Total Number`), position=position_dodge(width=0.9), vjust=-0.25)
reserved_room_chart
```



```
#actual_room type
actual_room <- table(Hotel_bookings$assigned_room_type)
actual_room_df <- as.data.frame(actual_room)
colnames(actual_room_df) <- c('Actual room types', 'Total Number')
actual_room_df
```

```
##    Actual room types Total Number
## 1                  A        74053
## 2                  B         2163
## 3                  C         2375
## 4                  D        25322
## 5                  E         7806
## 6                  F         3751
## 7                  G         2553
## 8                  H          712
## 9                  I          363
## 10                 K          279
## 11                 L            1
## 12                 P           12
```

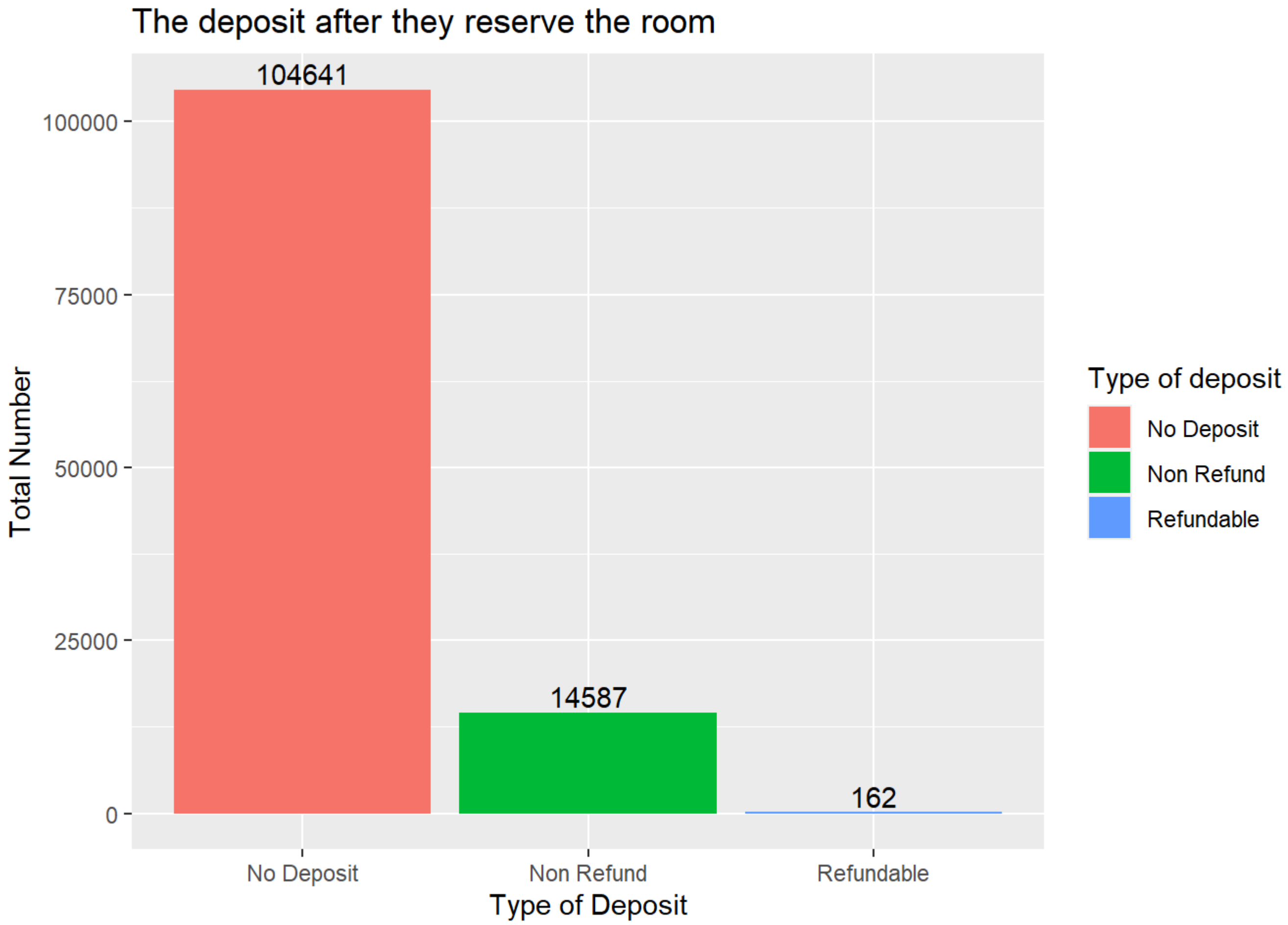
```
actual_room_chart <- ggplot(data=actual_room_df, aes(x=`Actual room types`, y=`Total Number`, fi
ll=`Actual room types`))+
  geom_bar(stat="identity")+
    labs(title = "Type of room that the guests actually received", fill = "Type of room")+
  geom_text(aes(label=`Total Number`), position=position_dodge(width=0.9), vjust=-0.25)
actual_room_chart
```



```
#deposit types
deposit <- table(Hotel_bookings$deposit_type)
deposit_df <- as.data.frame(deposit)
colnames(deposit_df) <- c('Type of Deposit', 'Total Number')
deposit_df
```

```
##   Type of Deposit Total Number
## 1      No Deposit       104641
## 2      Non Refund        14587
## 3      Refundable          162
```

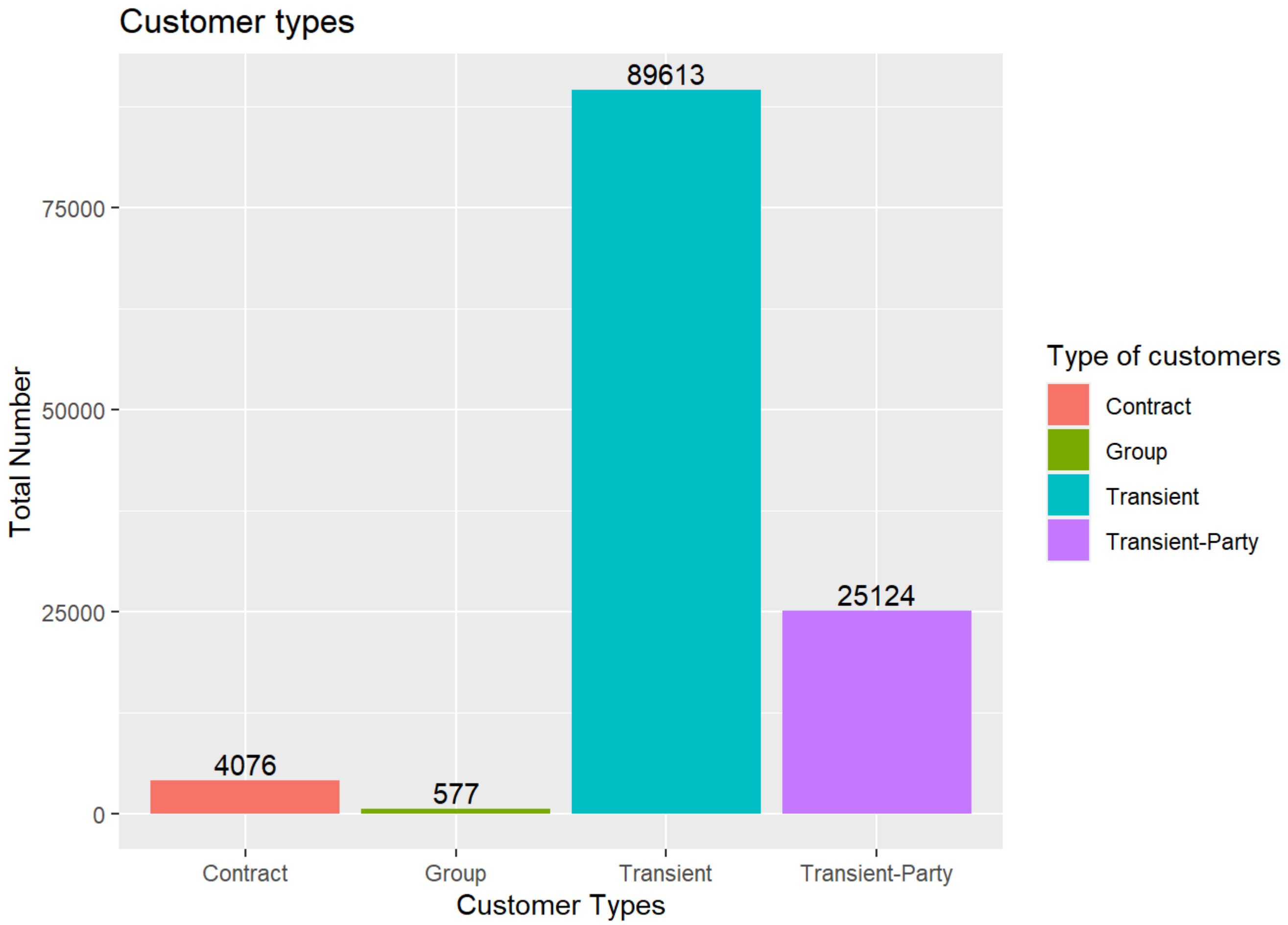
```
# Create the pie chart for market segment
deposit_chart<- ggplot(data=deposit_df, aes(x=`Type of Deposit`, y=`Total Number`, fill=`Type of
Deposit`))+
  geom_bar(stat="identity")+
    labs(title = "The deposit after they reserve the room", fill = "Type of deposit")+
  geom_text(aes(label=`Total Number`), position=position_dodge(width=0.9), vjust=-0.25)
deposit_chart
```



```
#Customer types
customer <- table(Hotel_bookings$customer_type)
customer_df <- as.data.frame(customer)
colnames(customer_df) <- c('Customer Types', 'Total Number')
customer_df
```

```
##    Customer Types Total Number
## 1        Contract         4076
## 2           Group          577
## 3       Transient        89613
## 4 Transient-Party        25124
```

```
customer_chart<- ggplot(data=customer_df, aes(x=`Customer Types`, y=`Total Number`, fill=`Custom
er Types`))+
  geom_bar(stat="identity")+
    labs(title = "Customer types", fill = "Type of customers")+
  geom_text(aes(label=`Total Number`), position=position_dodge(width=0.9), vjust=-0.25)
customer_chart
```



```
#Create the country distribution
country <- table(Hotel_bookings$country)
country_df <- as.data.frame(country)
head(country_df)
```

```
##   Var1 Freq
## 1  ABW    2
## 2  AGO  362
## 3  AIA    1
## 4  ALB   12
## 5  AND    7
## 6  ARE   51
```

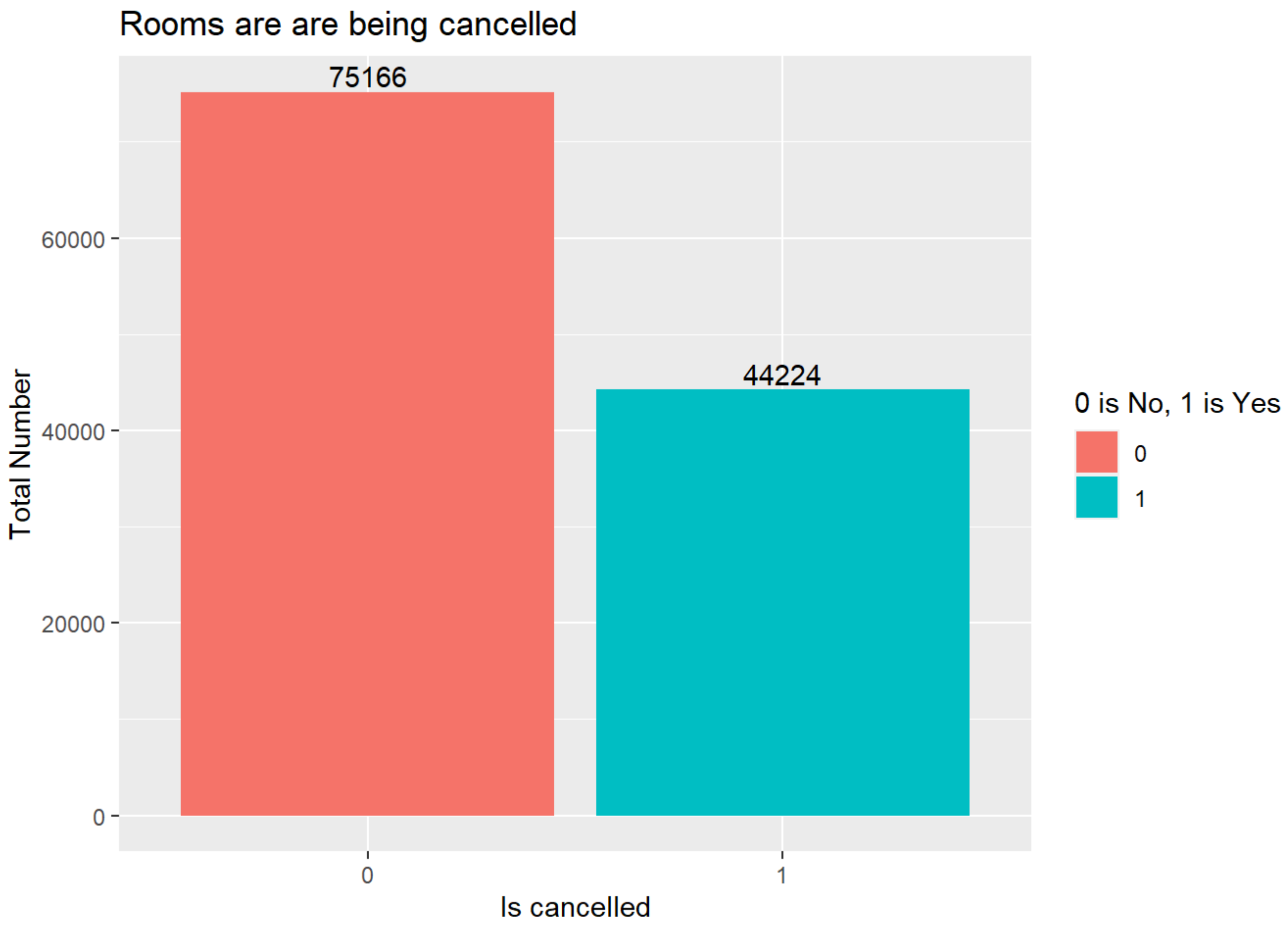
```
#Matched room
being_cancelled <- table(Hotel_bookings$is_canceled)
being_cancelled_df <- as.data.frame(being_cancelled)
colnames(being_cancelled_df) <- c('Is cancelled', 'Total Number')
being_cancelled_df
```

```
##   Is cancelled Total Number
## 1            0        75166
## 2            1        44224
```

```
head(as.character(Hotel_bookings$is_canceled))
```

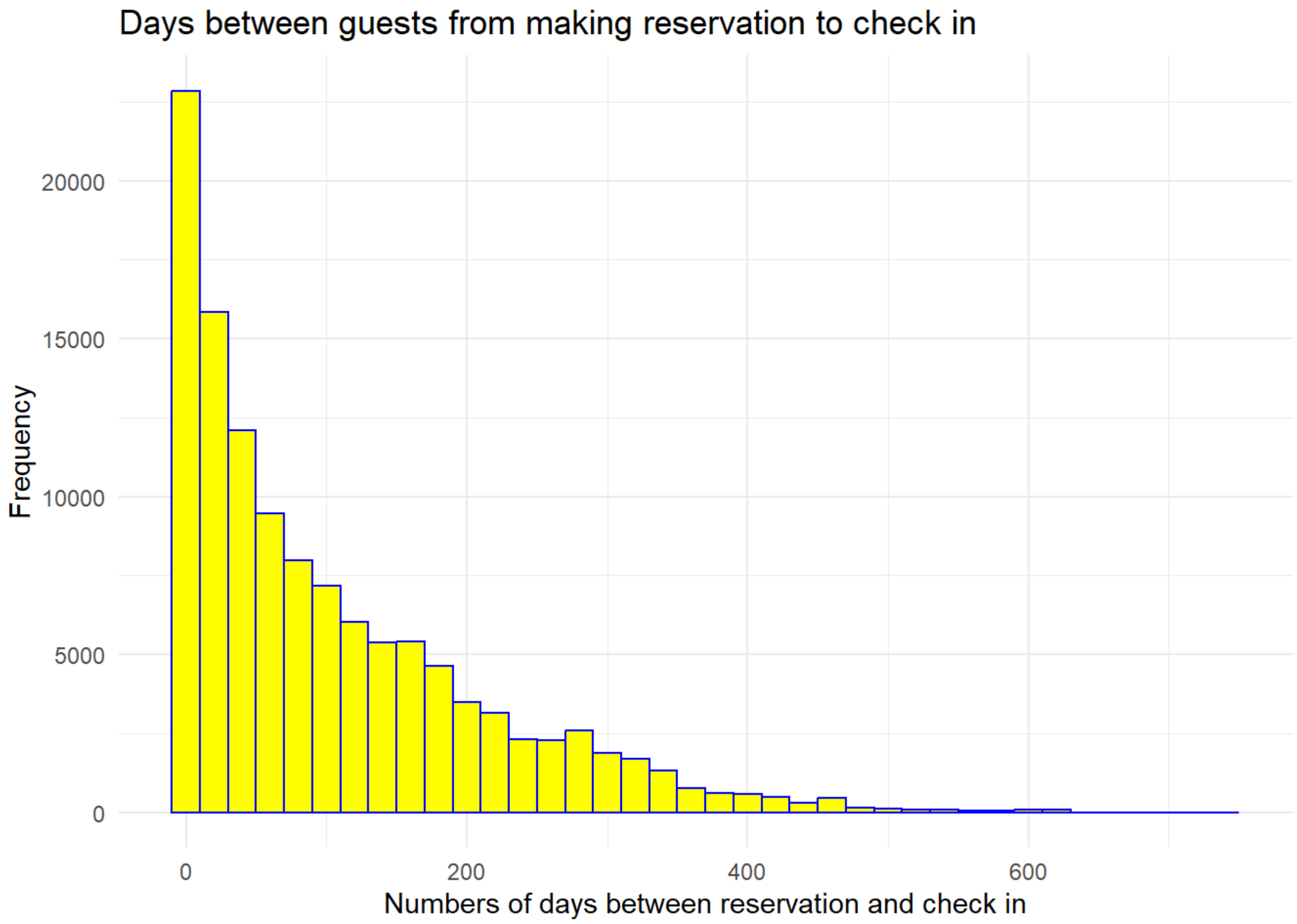
```
## [1] "0" "0" "0" "0" "0" "0"
```

```
# Create the pie chart for market segment
being_cancelled_chart<- ggplot(data=being_cancelled_df, aes(x=`Is cancelled`, y=`Total Number`,
fill=`Is cancelled`))+
  geom_bar(stat="identity")+
    labs(title = "Rooms are are being cancelled", fill = "0 is No, 1 is Yes")+
  geom_text(aes(label=`Total Number`), position=position_dodge(width=0.9), vjust=-0.25)
being_cancelled_chart
```



```
#1 is being cancelled, 0 is not cancelled
```

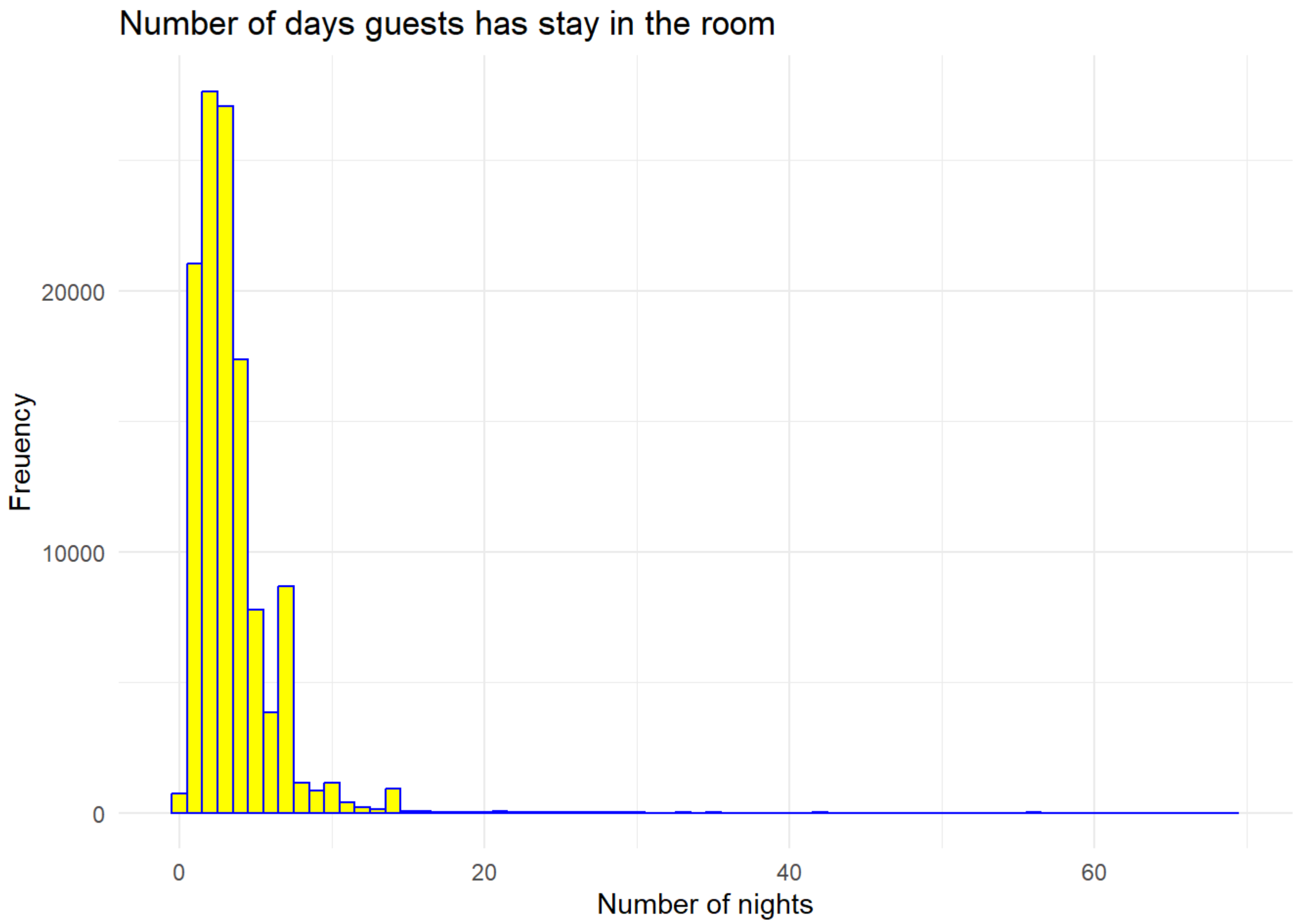
```
#EDA for numerical portion
#lead_time distribution
ggplot(Hotel_bookings, aes(x = lead_time)) +
  geom_histogram(binwidth = 20, fill = "yellow", color = "Blue") +
  labs(
    title = "Days between guests from making reservation to check in",
    x = "Numbers of days between reservation and check in",
    y= "Frequency"
  ) +
  theme_minimal()
```



```
#number of nights stayed
Hotel_bookings$stays <- Hotel_bookings$stays_in_weekend_nights+Hotel_bookings$stays_in_week_nigh
ts
head(as.numeric(Hotel_bookings$stays))
```

```
## [1] 0 0 1 1 2 2
```

```
ggplot(Hotel_bookings, aes(x = stays)) +
  geom_histogram(binwidth = 1, fill = "yellow", color = "Blue") +
  labs(
    title = "Number of days guests has stay in the room",
    x = "Number of nights",
    y = "Freuency"
  ) +
  theme_minimal()
```

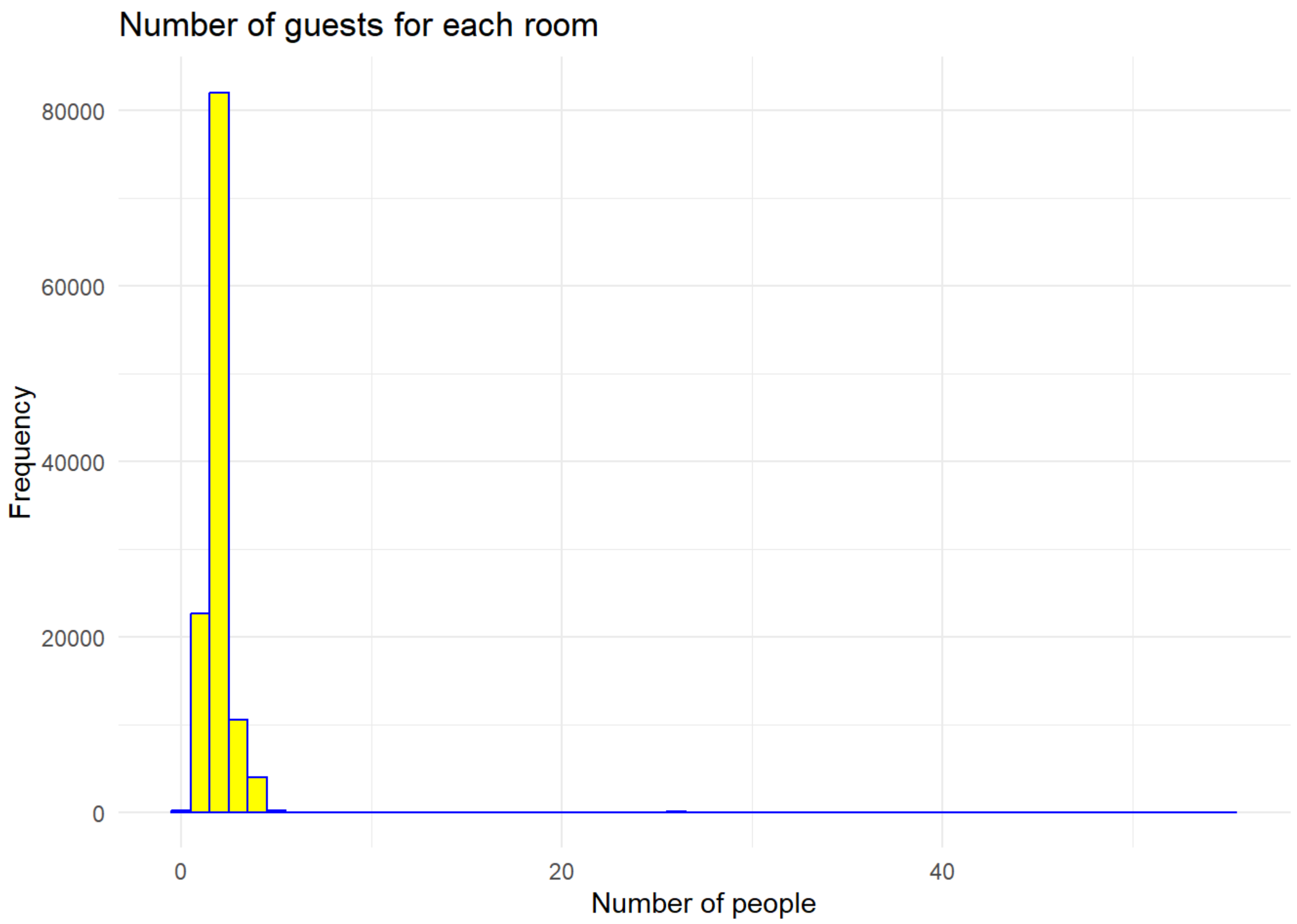


```
#number of total person

Hotel_bookings$people <- Hotel_bookings$adults+Hotel_bookings$children+Hotel_bookings$babies
head(as.numeric(Hotel_bookings$people))
```

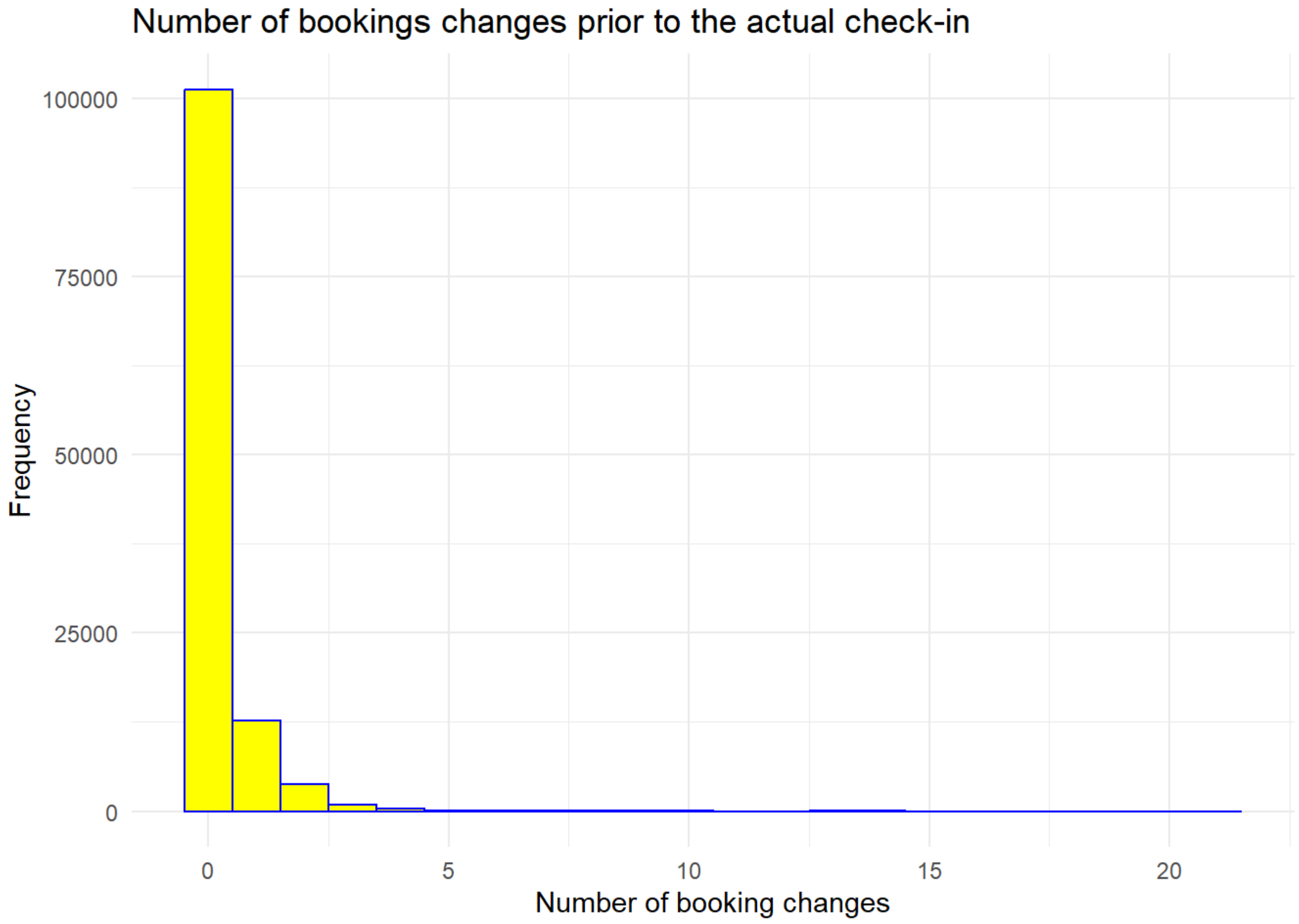
```
## [1] 2 2 1 1 2 2
```

```
ggplot(Hotel_bookings, aes(x = people)) +
  geom_histogram(binwidth = 1, fill = "yellow", color = "Blue") +
  labs(
    title = "Number of guests for each room",
    x = "Number of people",
    y = "Frequency"
  ) +
  theme_minimal()
```

```
## Warning: Removed 4 rows containing non-finite values (`stat_bin()`).
```

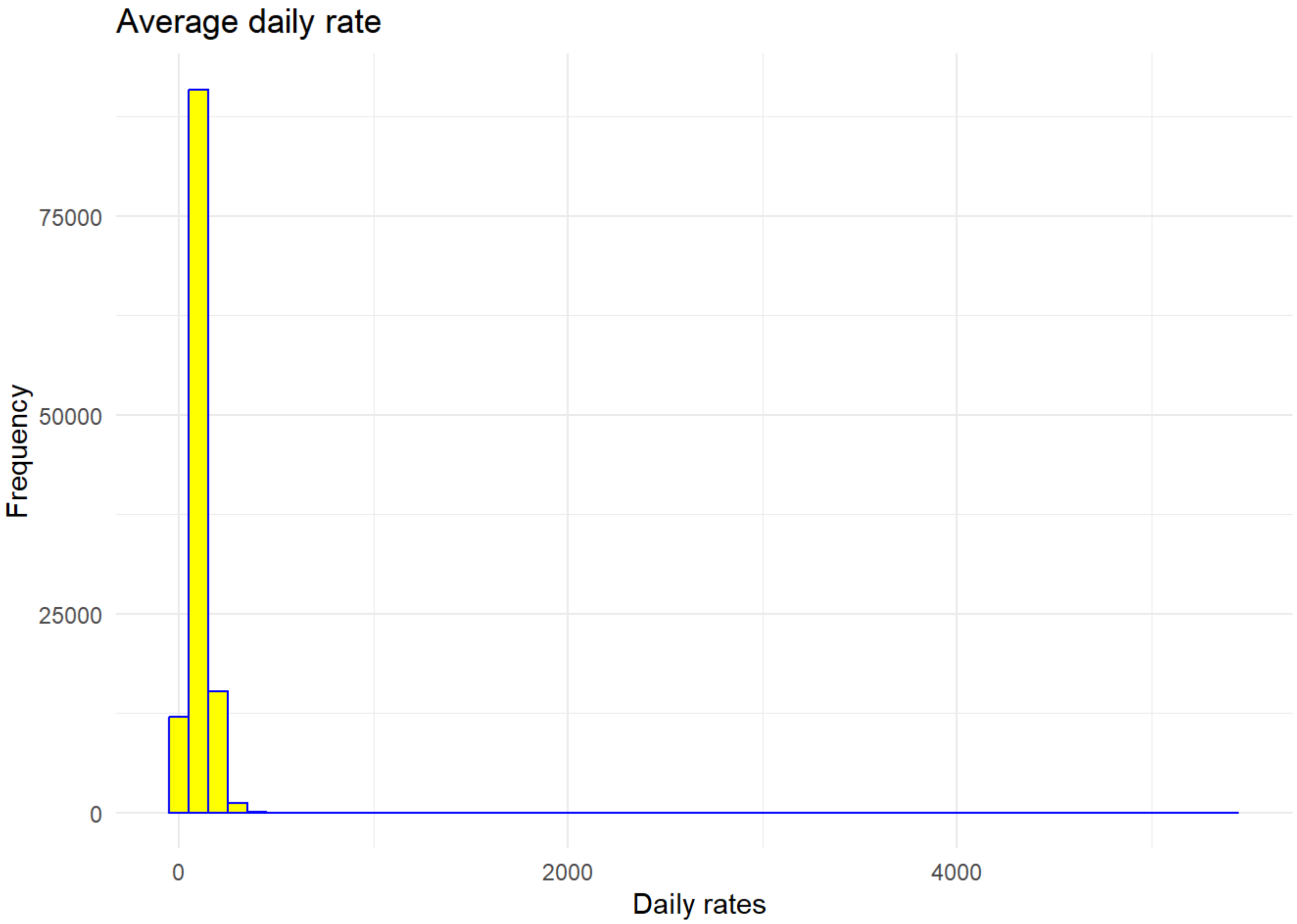


```
#booking changes
ggplot(Hotel_bookings, aes(x = booking_changes)) +
  geom_histogram(binwidth = 1, fill = "yellow", color = "Blue") +
  labs(
    title = "Number of bookings changes prior to the actual check-in",
    x = "Number of booking changes",
    y = "Frequency"
  ) +
  theme_minimal()
```

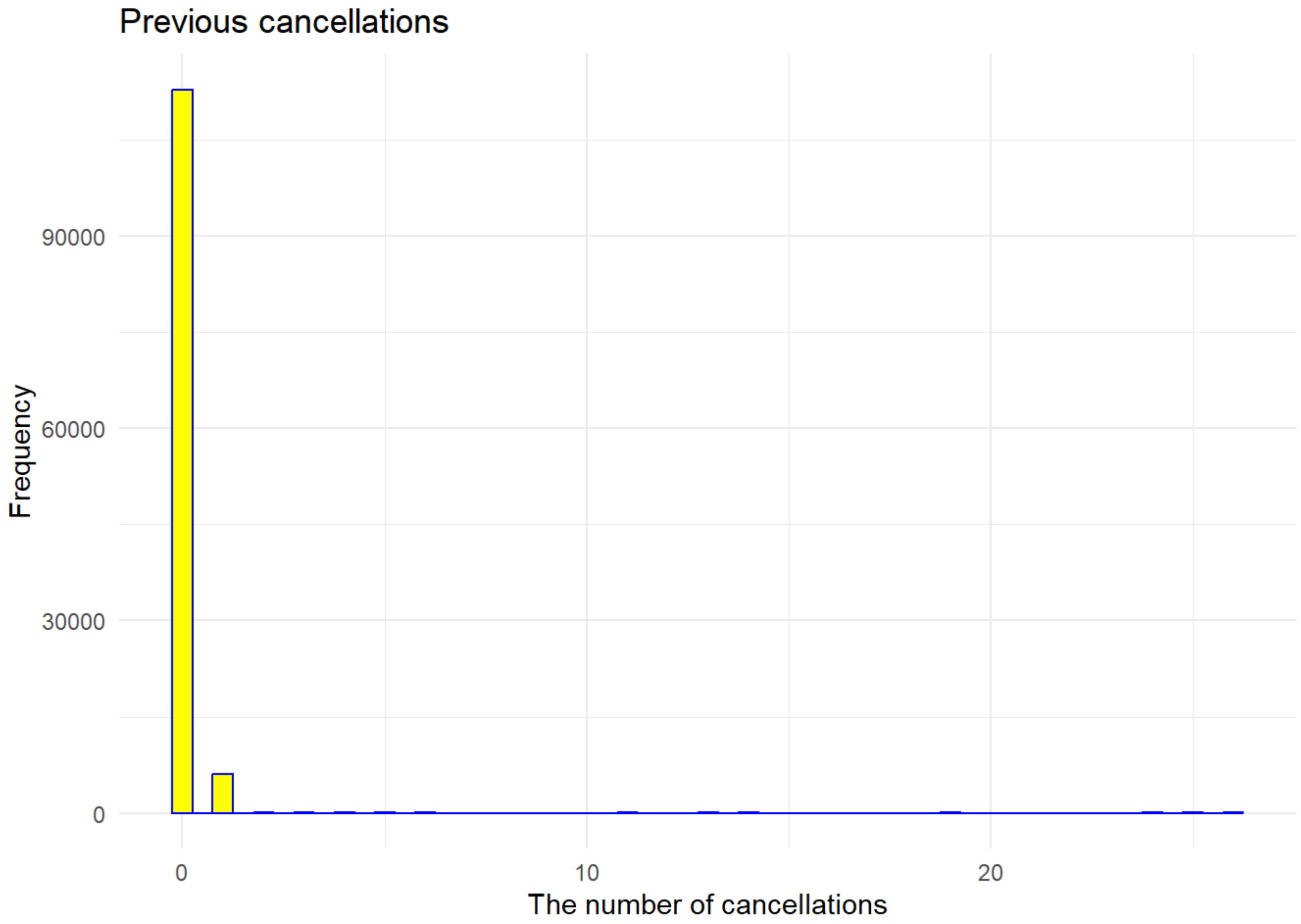


Number of bookings changes prior to the actual check-in

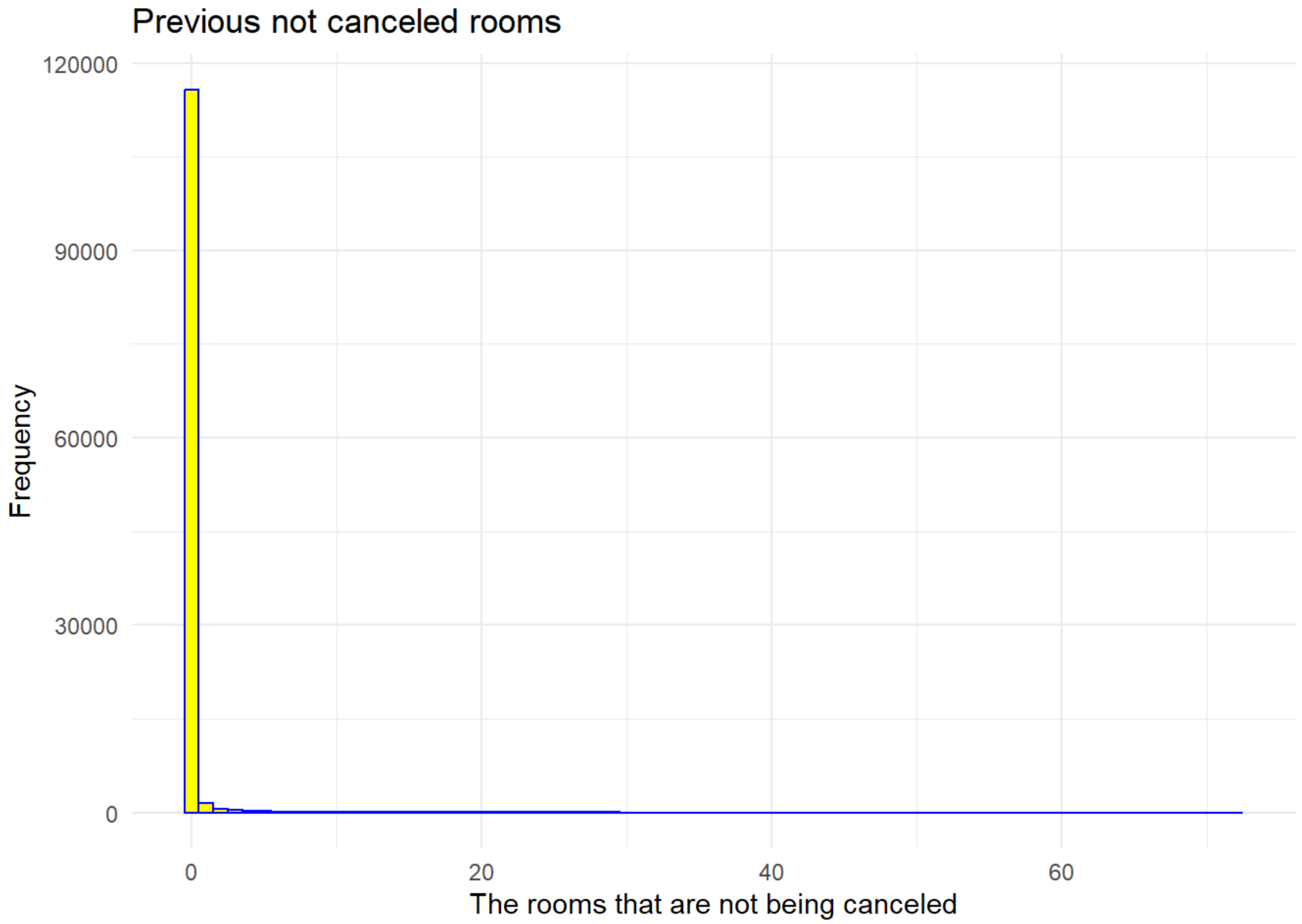
```
#adr
ggplot(Hotel_bookings, aes(x = adr)) +
  geom_histogram(binwidth = 100, fill = "yellow", color = "Blue") +
  labs(
    title = "Average daily rate",
    x = "Daily rates ",
    y = "Frequency"
  ) +
  theme_minimal()
```



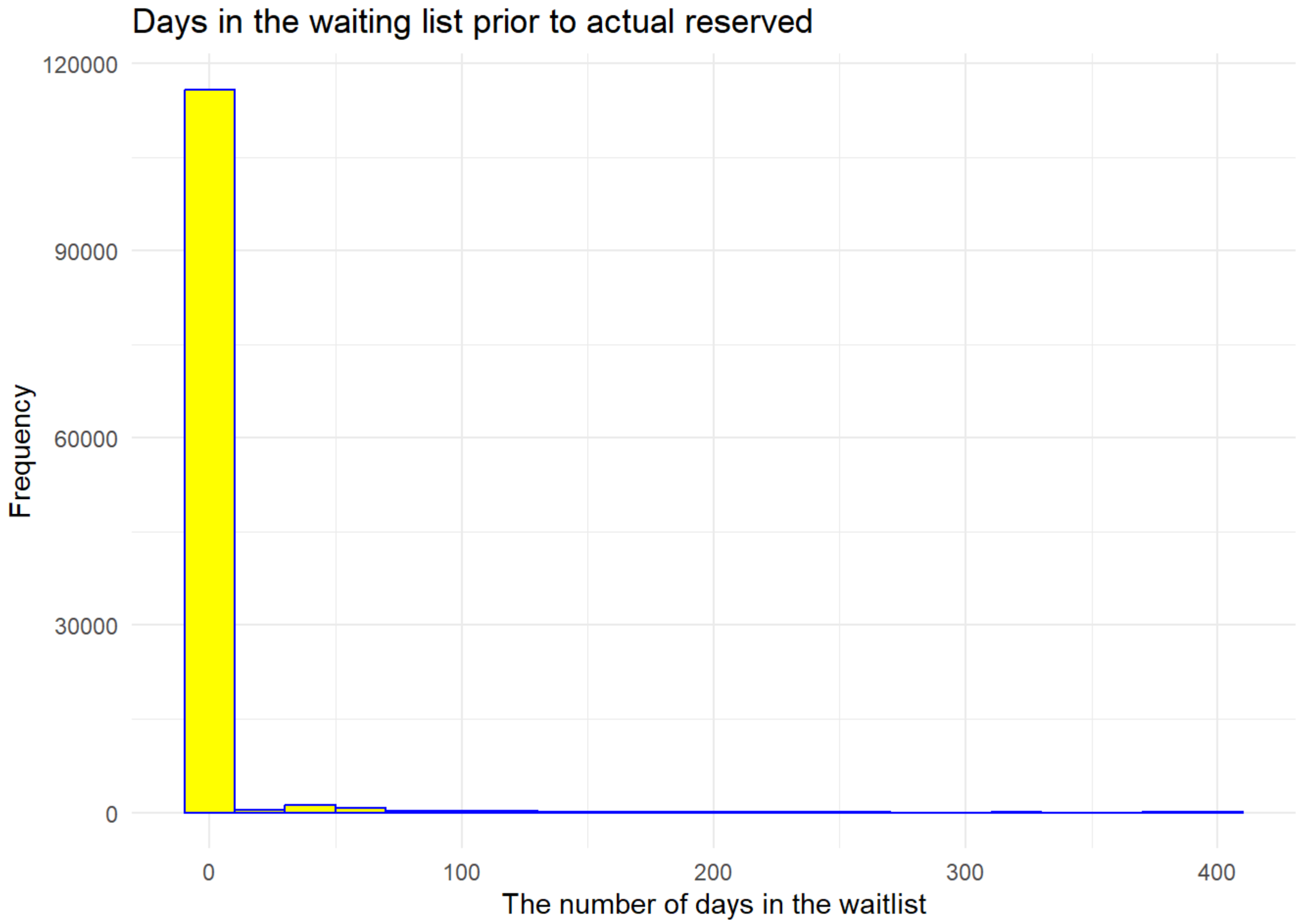
```
#previous cancellations
ggplot(Hotel_bookings, aes(x = previous_cancellations)) +
  geom_histogram(binwidth = 0.5, fill = "yellow", color = "Blue") +
  labs(
    title = "Previous cancellations",
    x = "The number of cancellations",
    y = "Frequency"
  ) +
  theme_minimal()
```



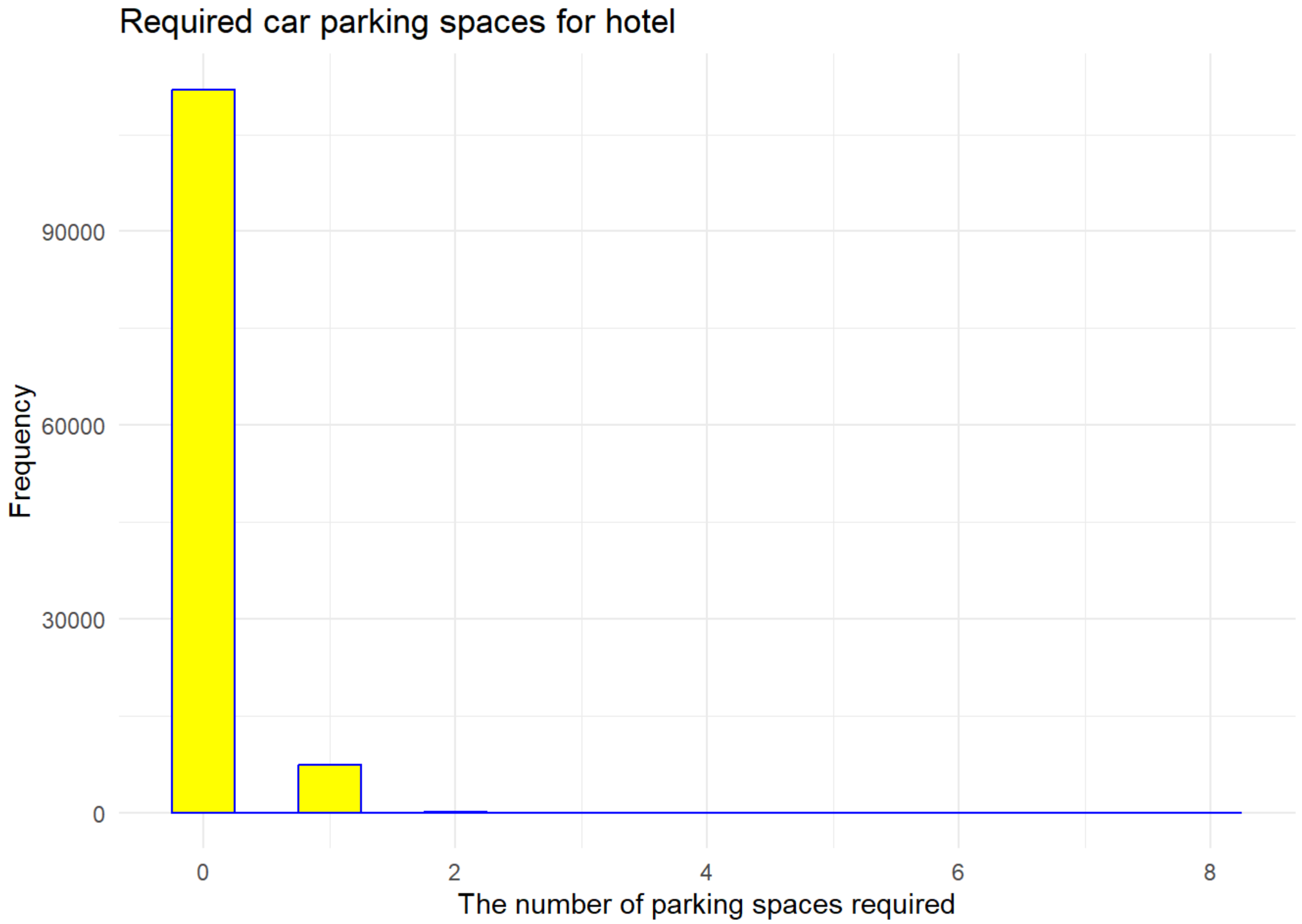
```
#previous not canceled room
ggplot(Hotel_bookings, aes(x = previous_bookings_not_canceled)) +
  geom_histogram(binwidth = 1, fill = "yellow", color = "Blue") +
  labs(
    title = "Previous not canceled rooms",
    x = "The rooms that are not being canceled",
    y = "Frequency"
  ) +
  theme_minimal()
```



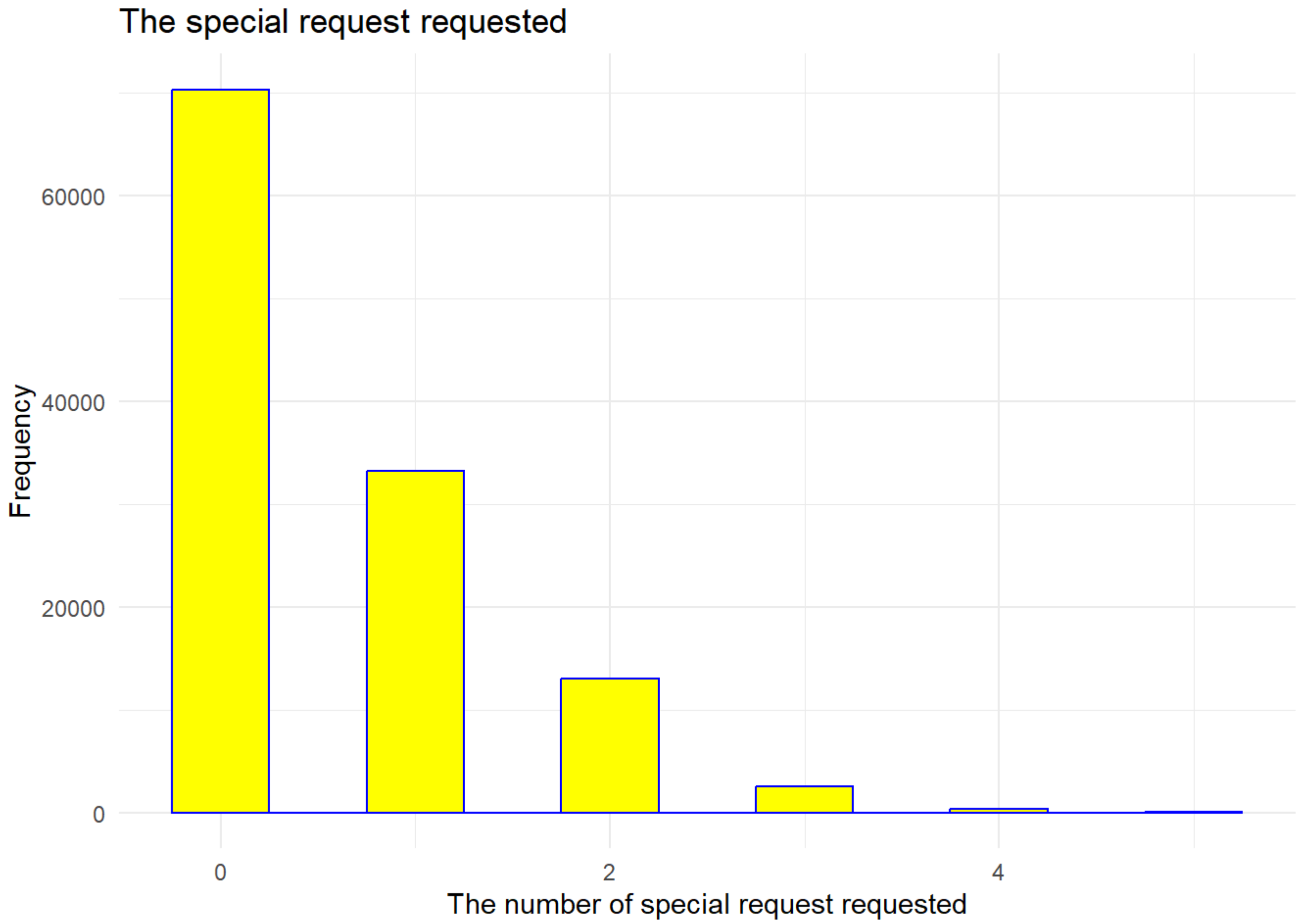
```
#Waiting list
ggplot(Hotel_bookings, aes(x = days_in_waiting_list)) +
  geom_histogram(binwidth = 20, fill = "yellow", color = "Blue") +
  labs(
    title = "Days in the waiting list prior to actual reserved",
    x = "The number of days in the waitlist",
    y = "Frequency"
  ) +
  theme_minimal()
```



```
#require spaces
ggplot(Hotel_bookings, aes(x = required_car_parking_spaces)) +
  geom_histogram(binwidth = 0.5, fill = "yellow", color = "Blue") +
  labs(
    title = "Required car parking spaces for hotel",
    x = "The number of parking spaces required",
    y = "Frequency"
  ) +
  theme_minimal()
```



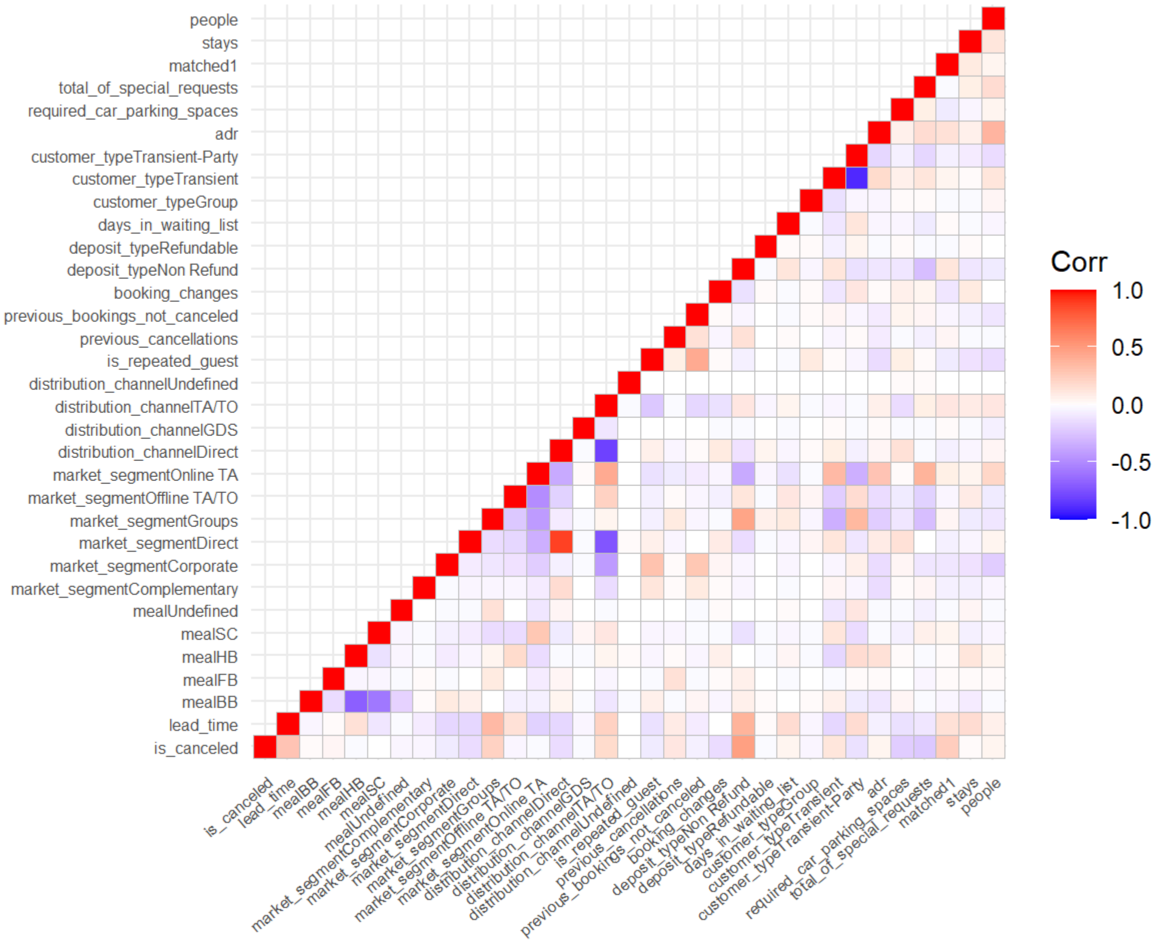
```
#special requests
ggplot(Hotel_bookings, aes(x = total_of_special_requests)) +
  geom_histogram(binwidth = 0.5, fill = "yellow", color = "Blue") +
  labs(
    title = "The special request requested",
    x = "The number of special request requested",
    y = "Frequency"
  ) +
  theme_minimal()
```



```
#delete unnecessary columns
Hotel_bookings2 <- subset(Hotel_bookings, select = -c(arrival_date_year,arrival_date_month, arri
val_date_week_number, arrival_date_day_of_month, agent, company, reservation_status_date, adult
s, children, babies, stays_in_weekend_nights, stays_in_week_nights, country, hotel, reservation_
status, reserved_room_type, assigned_room_type))
head(Hotel_bookings2)
```

```
## # A tibble: 6 x 18
##   is_canceled lead_time meal  market_segment distribution_channel
##         <dbl>     <dbl> <chr> <chr>          <chr>
## 1           0       342 BB    Direct         Direct
## 2           0       737 BB    Direct         Direct
## 3           0         7 BB    Direct         Direct
## 4           0        13 BB    Corporate      Corporate
## 5           0        14 BB    Online TA      TA/TO
## 6           0        14 BB    Online TA      TA/TO
## # i 13 more variables: is_repeated_guest <dbl>, previous_cancellations <dbl>,
## #   previous_bookings_not_canceled <dbl>, booking_changes <dbl>,
## #   deposit_type <chr>, days_in_waiting_list <dbl>, customer_type <chr>,
## #   adr <dbl>, required_car_parking_spaces <dbl>,
## #   total_of_special_requests <dbl>, matched <chr>, stays <dbl>, people <dbl>
```

```
#correlation matrix:

matrix<- model.matrix(~0+., data=Hotel_bookings2) %>%
  cor(use="pairwise.complete.obs") %>%
  ggcorrplot(show.diag=TRUE, type="lower", lab=FALSE, lab_size=1, tl.cex=6, tl.srt=40)

matrix
```



```
#Build out regression models: Full models
full_model <- glm(is_canceled ~., data=Hotel_bookings2, family=binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(full_model)
```

```
## 
## Call:
## glm(formula = is_canceled ~ ., family = binomial, data = Hotel_bookings2)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -8.4904  -0.7444  -0.3047   0.2046   5.9435  
## 
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                    -4.129e+00  1.838e-01 -22.465  < 2e-16 ***
## lead_time                       3.579e-03  9.309e-05  38.445  < 2e-16 ***
## mealFB                          7.938e-01  1.083e-01   7.331 2.28e-13 ***
## mealHB                         -8.222e-02  2.647e-02  -3.106 0.001894 ** 
## mealSC                          5.882e-02  2.459e-02   2.392 0.016745 *  
## mealUndefined                  -4.678e-01  9.857e-02  -4.746 2.07e-06 ***
## market_segmentComplementary     7.987e-01  2.254e-01   3.544 0.000395 ***
## market_segmentCorporate         9.784e-03  1.765e-01   0.055 0.955789    
## market_segmentDirect            2.113e-01  1.960e-01   1.078 0.281083    
## market_segmentGroups            2.444e-01  1.847e-01   1.324 0.185599    
## market_segmentOffline TA/TO    -3.656e-01  1.852e-01  -1.975 0.048306 *  
## market_segmentOnline TA         9.168e-01  1.845e-01   4.968 6.76e-07 ***
## distribution_channelDirect     -5.964e-01  9.542e-02  -6.251 4.09e-10 ***
## distribution_channelGDS        -1.161e+00  2.018e-01  -5.755 8.67e-09 ***
## distribution_channelTA/TO      -1.870e-01  7.108e-02  -2.631 0.008516 ** 
## distribution_channelUndefined   1.941e+03  7.673e+05   0.003 0.997981    
## is_repeated_guest              -6.213e-01  8.553e-02  -7.264 3.75e-13 ***
## previous_cancellations          2.724e+00  6.051e-02  45.019  < 2e-16 ***
## previous_bookings_not_canceled -4.914e-01  2.526e-02 -19.452  < 2e-16 ***
## booking_changes                -3.421e-01  1.524e-02 -22.456  < 2e-16 ***
## deposit_typeNon Refund          5.429e+00  1.127e-01  48.151  < 2e-16 ***
## deposit_typeRefundable          1.457e-01  2.149e-01   0.678 0.497738    
## days_in_waiting_list           -1.653e-04  4.812e-04  -0.344 0.731189    
## customer_typeGroup             -1.212e-01  1.713e-01  -0.707 0.479324    
## customer_typeTransient          8.585e-01  5.356e-02  16.031  < 2e-16 ***
## customer_typeTransient-Party    3.931e-01  5.699e-02   6.897 5.30e-12 ***
## adr                             3.230e-03  1.959e-04  16.486  < 2e-16 ***
## required_car_parking_spaces    -1.953e+03  7.673e+05  -0.003 0.997969    
## total_of_special_requests      -7.086e-01  1.152e-02 -61.488  < 2e-16 ***
## matched1                        1.778e+00  4.031e-02  44.101  < 2e-16 ***
## stays                           4.009e-02  3.128e-03  12.817  < 2e-16 ***
## people                          1.237e-01  1.281e-02   9.655  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 157390  on 119385  degrees of freedom
## Residual deviance:  99685  on 119354  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 99749
```

```
##
## Number of Fisher Scoring iterations: 12
```

```
anova(full_model)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Analysis of Deviance Table
## 
## Model: binomial, link: logit
## 
## Response: is_canceled
## 
## Terms added sequentially (first to last)
## 
## 
##                                Df Deviance Resid. Df Resid. Dev
## NULL                                          119385     157390
## lead_time                       1  10244.0    119384     147146
## meal                            4    768.2    119380     146378
## market_segment                  6   4145.3    119374     142233
## distribution_channel            4    471.7    119370     141761
## is_repeated_guest               1    191.2    119369     141570
## previous_cancellations          1   4419.8    119368     137150
## previous_bookings_not_canceled  1   1677.3    119367     135473
## booking_changes                 1   2471.7    119366     133001
## deposit_type                    2  19646.4    119364     113355
## days_in_waiting_list            1      0.7    119363     113354
## customer_type                   3    713.0    119360     112641
## adr                             1    525.5    119359     112115
## required_car_parking_spaces     1   4598.4    119358     107517
## total_of_special_requests       1   4529.0    119357     102988
## matched                         1   3022.4    119356      99966
## stays                           1    176.1    119355      99789
## people                          1    104.8    119354      99685
```

```
vif(full_model)
```

```
##                                        GVIF Df GVIF^(1/(2*Df))
## lead_time                      1.298135e+00  1        1.139357
## meal                           1.377405e+00  4        1.040837
## market_segment                 6.903104e+01  6        1.423160
## distribution_channel           5.170651e+07  4        9.208590
## is_repeated_guest              1.325286e+00  1        1.151211
## previous_cancellations         1.545963e+00  1        1.243367
## previous_bookings_not_canceled 1.624514e+00  1        1.274564
## booking_changes                1.034910e+00  1        1.017305
## deposit_type                   1.082540e+00  2        1.020025
## days_in_waiting_list           1.072591e+00  1        1.035660
## customer_type                  2.209880e+00  3        1.141287
## adr                            1.475681e+00  1        1.214776
## required_car_parking_spaces    2.053906e+06  1     1433.145343
## total_of_special_requests      1.184319e+00  1        1.088264
## matched                        1.016251e+00  1        1.008093
## stays                          1.158580e+00  1        1.076374
## people                         1.314950e+00  1        1.146713
```

```
#reduced model, after displayed the full model, key factors exposed to hotel cancellation

reduced_model <- glm(is_canceled ~ lead_time + meal +
    is_repeated_guest + previous_cancellations + previous_bookings_not_canceled +
    booking_changes + customer_type +
    adr + total_of_special_requests +
    stays + people + matched, data=Hotel_bookings2, family=binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(reduced_model)
```

```
##
## Call:
## glm(formula = is_canceled ~ lead_time + meal + is_repeated_guest +
##     previous_cancellations + previous_bookings_not_canceled +
##     booking_changes + customer_type + adr + total_of_special_requests +
##     stays + people + matched, family = binomial, data = Hotel_bookings2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.4904  -0.8436  -0.3956   0.8898   6.4027
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -4.202e+00  6.808e-02 -61.715  < 2e-16 ***
## lead_time                       5.956e-03  7.705e-05  77.293  < 2e-16 ***
## mealFB                          8.563e-01  8.741e-02   9.796  < 2e-16 ***
## mealHB                         -2.216e-01  2.330e-02  -9.510  < 2e-16 ***
## mealSC                          1.022e-01  2.367e-02   4.317 1.58e-05 ***
## mealUndefined                  -3.287e-01  8.238e-02  -3.990 6.60e-05 ***
## is_repeated_guest              -1.182e+00  8.364e-02 -14.133  < 2e-16 ***
## previous_cancellations          3.104e+00  5.690e-02  54.550  < 2e-16 ***
## previous_bookings_not_canceled -6.041e-01  2.617e-02 -23.085  < 2e-16 ***
## booking_changes                -5.239e-01  1.550e-02 -33.790  < 2e-16 ***
## customer_typeGroup             -2.166e-02  1.640e-01  -0.132 0.894950
## customer_typeTransient          1.484e+00  5.229e-02  28.372  < 2e-16 ***
## customer_typeTransient-Party    2.029e-01  5.462e-02   3.714 0.000204 ***
## adr                             3.569e-03  1.676e-04  21.301  < 2e-16 ***
## total_of_special_requests      -7.997e-01  1.061e-02 -75.370  < 2e-16 ***
## stays                          -1.142e-02  2.958e-03  -3.861 0.000113 ***
## people                          4.653e-03  1.039e-02   0.448 0.654263
## matched1                        2.089e+00  3.842e-02  54.363  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 157390  on 119385  degrees of freedom
## Residual deviance: 118986  on 119368  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 119022
##
## Number of Fisher Scoring iterations: 8
```

```
anova(reduced_model)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: algorithm did not converge
```

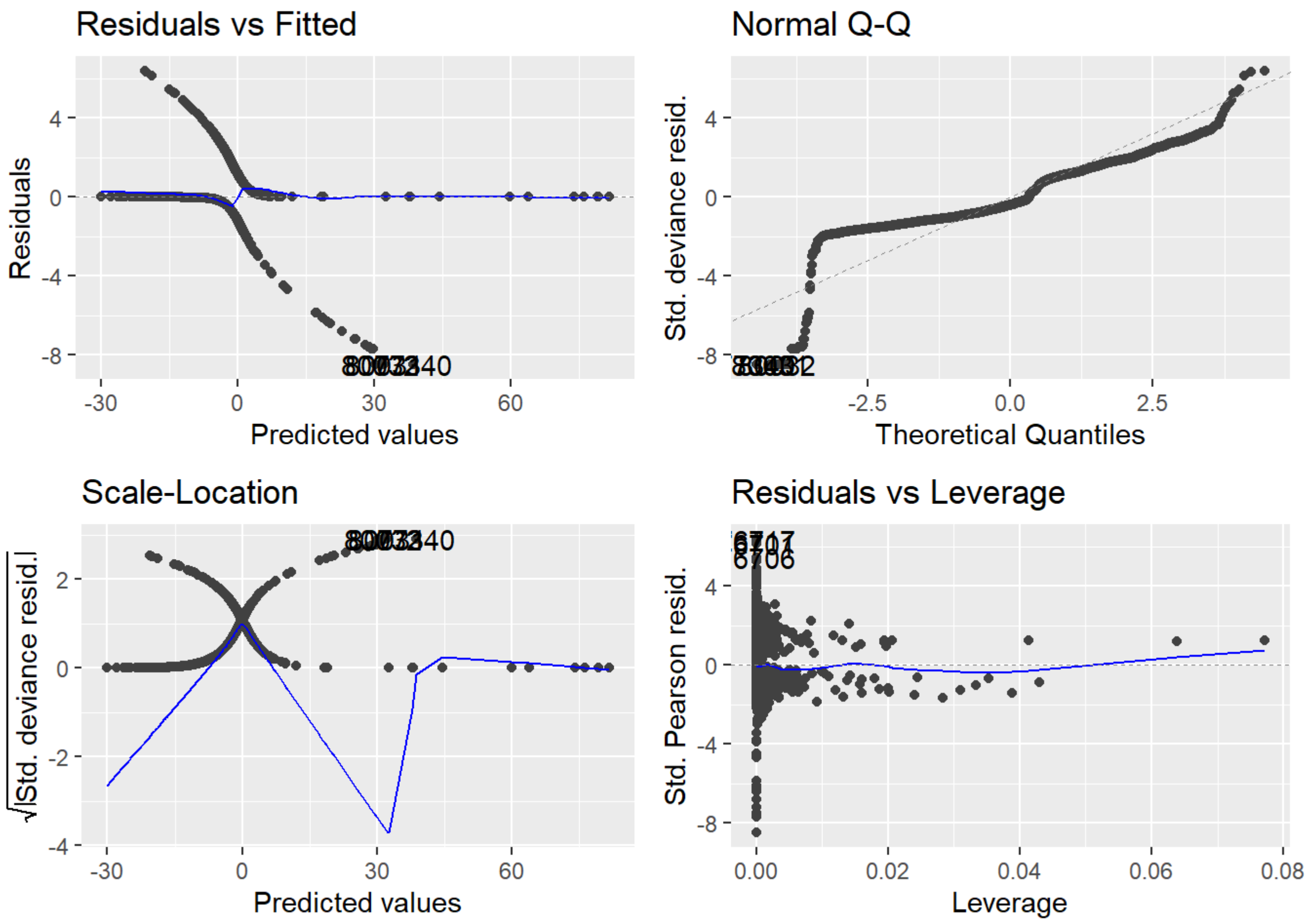
```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Analysis of Deviance Table
## 
## Model: binomial, link: logit
## 
## Response: is_canceled
## 
## Terms added sequentially (first to last)
## 
## 
##                                 Df Deviance Resid. Df Resid. Dev
## NULL                                          119385     157390
## lead_time                       1  10244.0    119384     147146
## meal                            4    768.2    119380     146378
## is_repeated_guest               1    410.9    119379     145967
## previous_cancellations          1   4946.4    119378     141021
## previous_bookings_not_canceled  1   2020.3    119377     139001
## booking_changes                 1   2738.8    119376     136262
## customer_type                   3   5041.3    119373     131220
## adr                             1    408.9    119372     130812
## total_of_special_requests       1   7022.6    119371     123789
## stays                           1      2.2    119370     123787
## people                          1      1.2    119369     123785
## matched                         1   4799.4    119368     118986
```

```
#Determine which one
anova(full_model, reduced_model)
```

```
## Analysis of Deviance Table
##
## Model 1: is_canceled ~ lead_time + meal + market_segment + distribution_channel +
##     is_repeated_guest + previous_cancellations + previous_bookings_not_canceled +
##     booking_changes + deposit_type + days_in_waiting_list + customer_type +
##     adr + required_car_parking_spaces + total_of_special_requests +
##     matched + stays + people
## Model 2: is_canceled ~ lead_time + meal + is_repeated_guest + previous_cancellations +
##     previous_bookings_not_canceled + booking_changes + customer_type +
##     adr + total_of_special_requests + stays + people + matched
##   Resid. Df Resid. Dev  Df Deviance
## 1    119354      99685
## 2    119368     118986 -14   -19301
```

```
#assumptions for the selection ones
autoplot(reduced_model)
```



```
#VIF for selected model
vif(reduced_model)
```

```
##                                    GVIF Df GVIF^(1/(2*Df))
## lead_time                      1.172163  1        1.082665
## meal                           1.180464  4        1.020955
## is_repeated_guest              1.285010  1        1.133583
## previous_cancellations         1.472305  1        1.213386
## previous_bookings_not_canceled 1.499041  1        1.224353
## booking_changes                1.020656  1        1.010275
## customer_type                  1.350050  3        1.051296
## adr                            1.278021  1        1.130496
## total_of_special_requests      1.072047  1        1.035397
## stays                          1.128518  1        1.062317
## people                         1.220434  1        1.104733
## matched                        1.013263  1        1.006609
```

```
durbinWatsonTest(reduced_model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.7600409     0.4799015       0
##  Alternative hypothesis: rho != 0
```

```
set.seed(1)
sample <- sample(c(TRUE, FALSE), nrow(Hotel_bookings2), replace=TRUE, prob=c(0.7,0.3))
train <- Hotel_bookings2[sample, ]
test <- Hotel_bookings2[!sample, ]

#AUC
prediction <- predict(reduced_model, test, type="response")
roc_object <- roc(test$is_canceled, prediction)
```

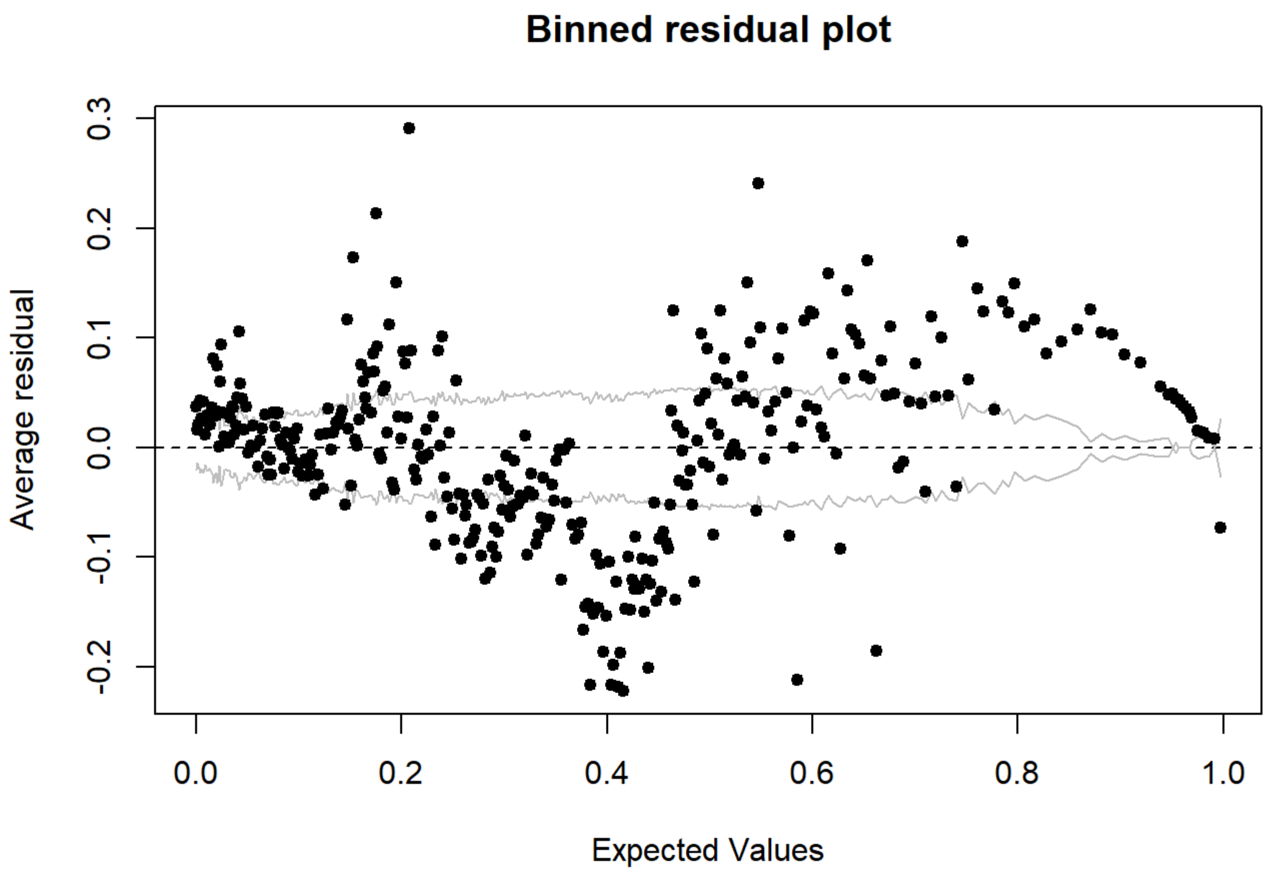
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

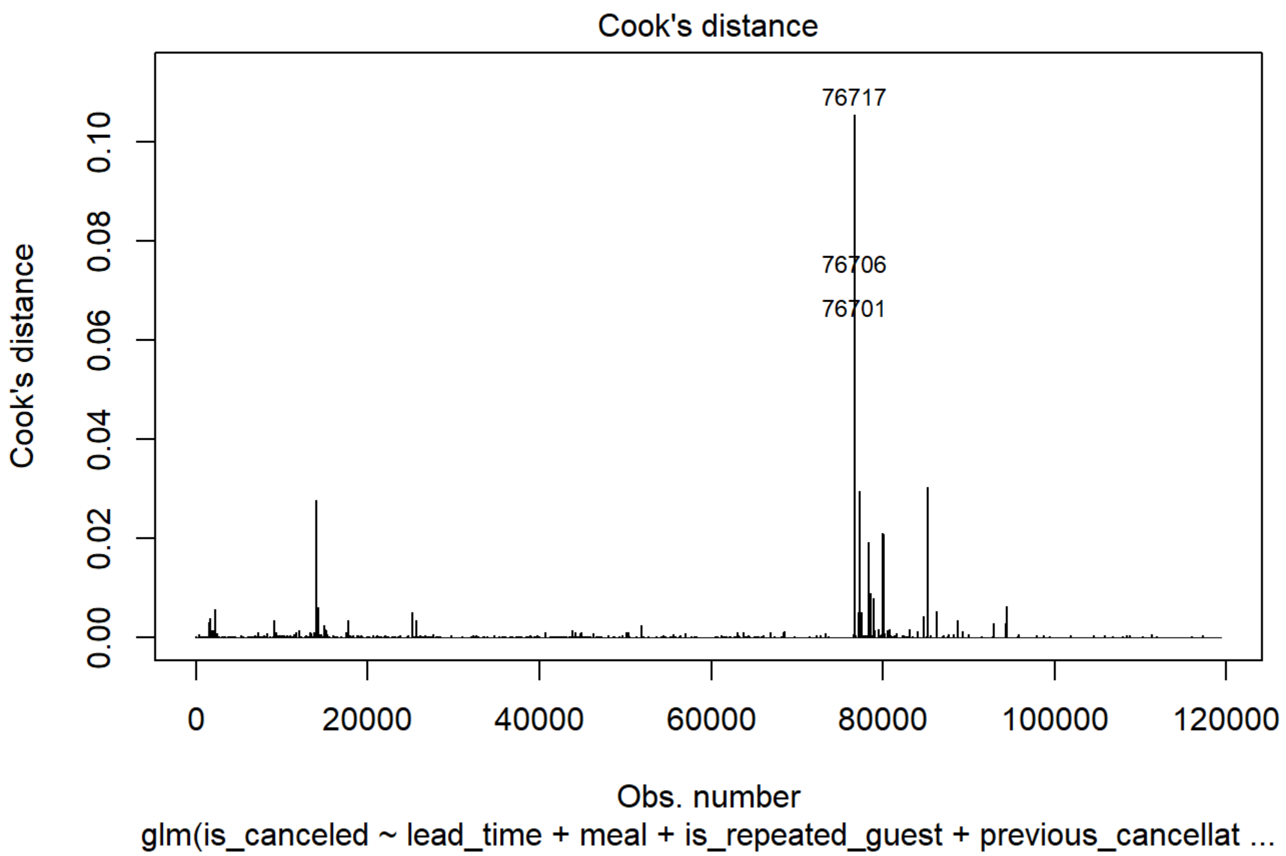
```
auc(roc_object)
```

```
## Area under the curve: 0.817
```

```
binnedplot(fitted(reduced_model),
           residuals(reduced_model, type="response"),
           nclass=NULL,
           xlab="Expected Values",
           ylab="Average residual",
           main="Binned residual plot",
           cex.pts=0.8,
           col.pts=1,
           col.int="gray")
```



```
plot(reduced_model, which = 4, id.n = 3)
```



```
anov <- aov(reduced_model)
anov
```

```
## Call:
##    aov(formula = reduced_model)
##
## Terms:
##                 lead_time      meal is_repeated_guest previous_cancellations
## Sum of Squares   2393.028   165.065            66.059                209.853
## Deg. of Freedom          1         4                 1                      1
##                 previous_bookings_not_canceled booking_changes customer_type
## Sum of Squares                          24.735         523.532       793.919
## Deg. of Freedom                              1               1             3
##                       adr total_of_special_requests     stays    people
## Sum of Squares     48.376                  1465.415     2.695     0.470
## Deg. of Freedom         1                         1         1         1
##                   matched Residuals
## Sum of Squares    774.652 21373.327
## Deg. of Freedom         1    119368
##
## Residual standard error: 0.4231478
## Estimated effects may be unbalanced
## 4 observations deleted due to missingness
```