Hotel Cancellation

Leo Shi

Why do I want to study hotel?

Every year, thousands of thousands of tourists travelling around the world to visit some of the famous sightseeing, but at the same time, they need a place to live.

Sometimes, booking a hotel could be a very frustrated, where numerous factors has to be considered, such as locations, prices, and services. At the same time, bookings could be suddenly got cancelled for a lot of reasons, such as room adjustment, marketing and so on. In 2018, around 40% of the room got cancelled even after being reserved.

Dataset introduction and goals

- There are two types of hotel: Resort hotel and city hotel.
- Conducted from July 2015 to August 2017
- I wanted to see that are the factors that would led their hotel room got cancelled after some room being reserved months ago.
- Initially, the dataset was containing 119390 rows of data and 32 variables
- Dataset was very clean enough, so there is no need to delete missing dataset
- But there is some adjustment on some of the dataset has been removed or being adjusted
- In the end, there are 18 variables remain for my project.

Variables introduction

- Hotel: the type of hotel that being booked (Resort or City)
- Is_canceled: indicate that the hotel was being cancelled: 1 is Cancelled, 0 is not
- Lead_time: number of days from reserved online to actual check-in or being cancelled
- Meal: type of meals that they previous booked (BB, FB, HB, SC)
- Market_segment: means the way of divide their market to the guests
- Distribution_channel: the way of the hotel distributed their room to the party that guest can able to book the room
- is_repeated_guests: Identify was a previously booked guests or not
- Matched: guests that has receive the same room that they initially reserved

Variables introduction continued

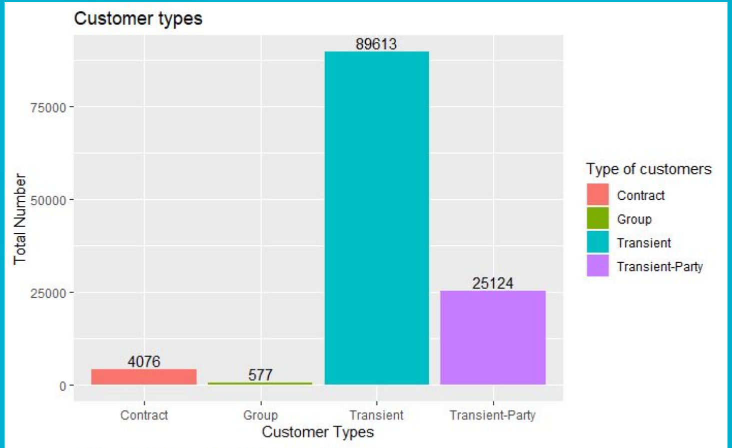
- Stays: Number of days that guests live
- People: Number of person within the group
- Previous_cancellations: Number of previous bookings that were cancelled by the customer prior to the current booking
- Previous_bookings_not_cancelled: Number of previous bookings not cancelled by the customer prior to the current booking
- Booking_changes: number of bookings changes prior to the final reservation
- Deposit_type: type of deposit prior to the booking
- Adr: Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
- Reservation_status: the status until now (Check-out or Cancelled)

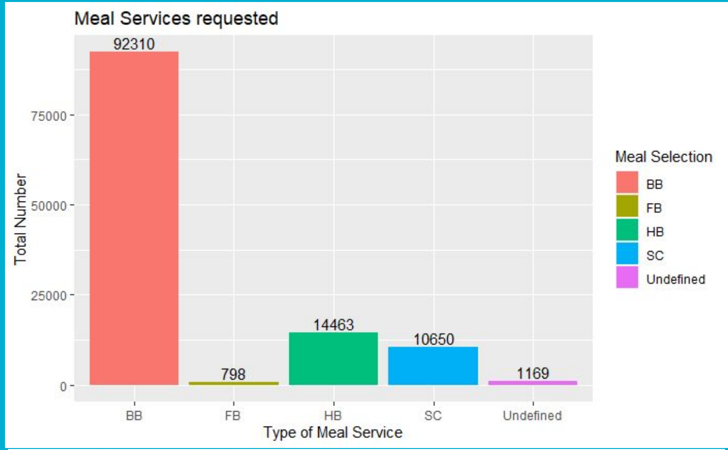
Variables introduction continued

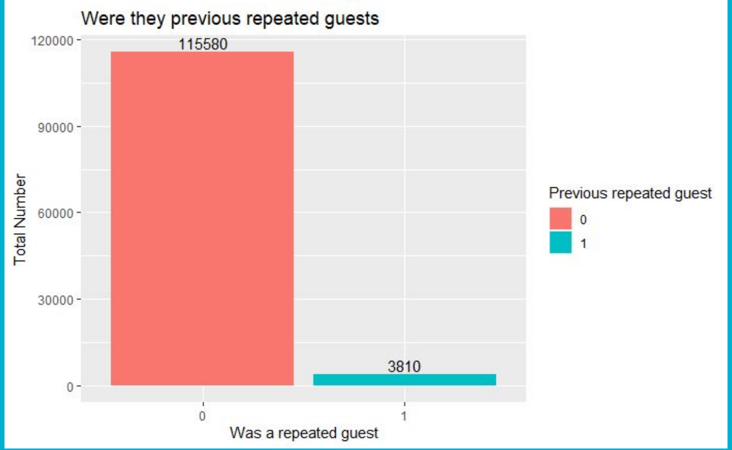
- Customer_type: the way that group of guests booking their hotel
- Required_car_parking_spaces: the number of parking slot need when guests made it to the hotel
- Total_of_special_requests: the extra request that request by the hotel, such as extra bed, higher bed, wheelchair service, etc.)

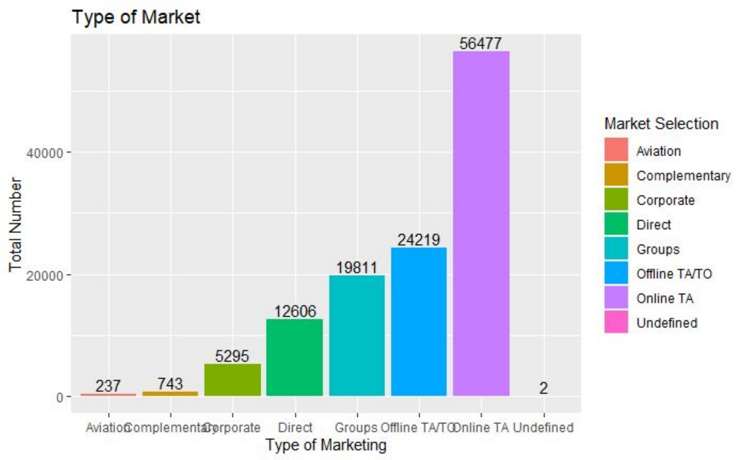
	is_canceled	lead_time	meal	market_segment	distribution_channel	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled	booking_changes	deposit_type	days_in_waiting_list	customer_type	adr	required_car_parking_spaces	total_of_special_requests
1	0	342	BB	Direct	Direct	0	0	0	3	No Deposit	0	Transient	0.00	0	0
2	0	737	BB	Direct	Direct	0	0	0	4	No Deposit	0	Transient	0.00	0	0
3	0	7	BB	Direct	Direct	0	0	0	0	No Deposit	0	Transient	75.00	0	0
4	0	13	BB	Corporate	Corporate	0	0	0	0	No Deposit	0	Transient	75.00	0	0
5	0	14	BB	Online TA	TA/TO	0	0	0	0	No Deposit	0	Transient	98.00	0	1
6	0	14	BB	Online TA	TA/TO	0	0	0	0	No Deposit	0	Transient	98.00	0	1
7	0	0	BB	Direct	Direct	0	0	0	0	No Deposit	0	Transient	107.00	0	0
8	0	9	FB	Direct	Direct	0	0	0	0	No Deposit	0	Transient	103.00	0	1
9	1	85	BB	Online TA	TA/TO	0	0	0	0	No Deposit	0	Transient	82.00	0	1
10	1	75	HB	Offline TA/TO	TA/TO	0	0	0	0	No Deposit	0	Transient	105.50	0	0
11	1	23	BB	Online TA	TA/TO	0	0	0	0	No Deposit	0	Transient	123.00	0	0
12	0	35	HB	Online TA	TA/TO	0	0	0	0	No Deposit	0	Transient	145.00	0	0
13	0	68	BB	Online TA	TA/TO	0	0	0	0	No Deposit	0	Transient	97.00	0	3

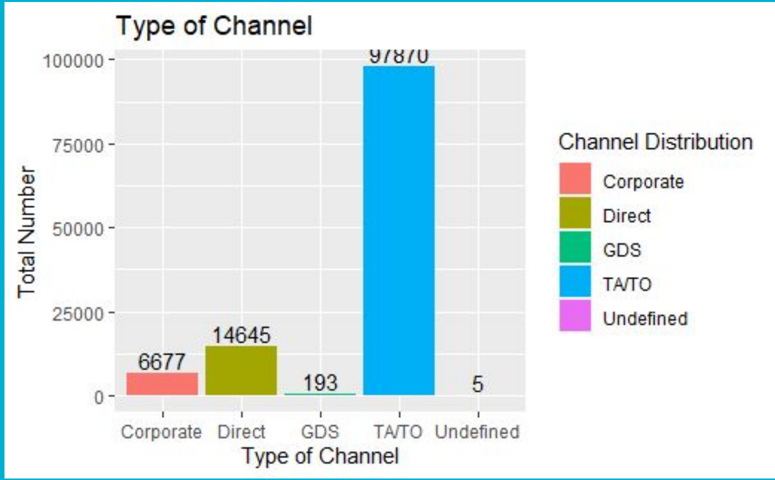
Exploratory Data Analysis: Categorical

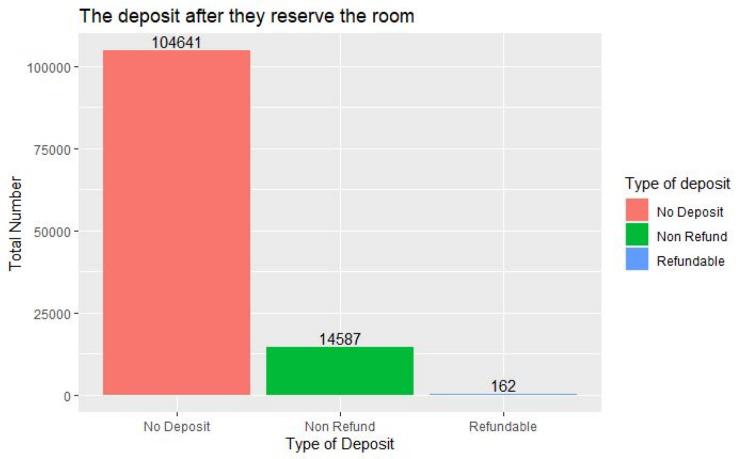


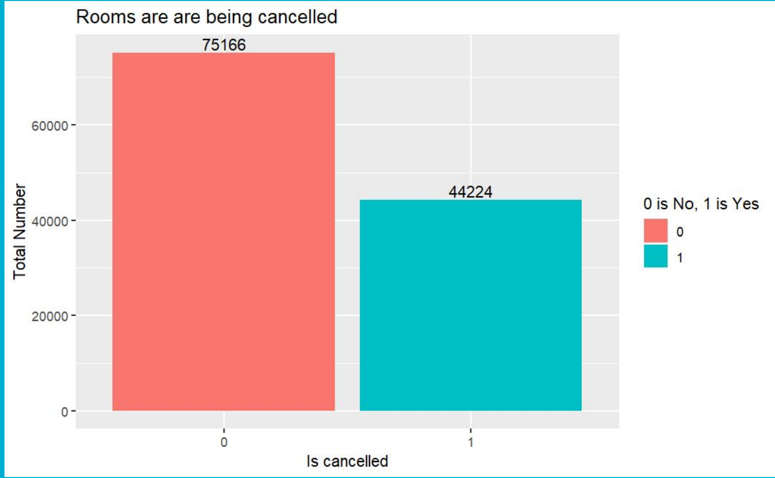




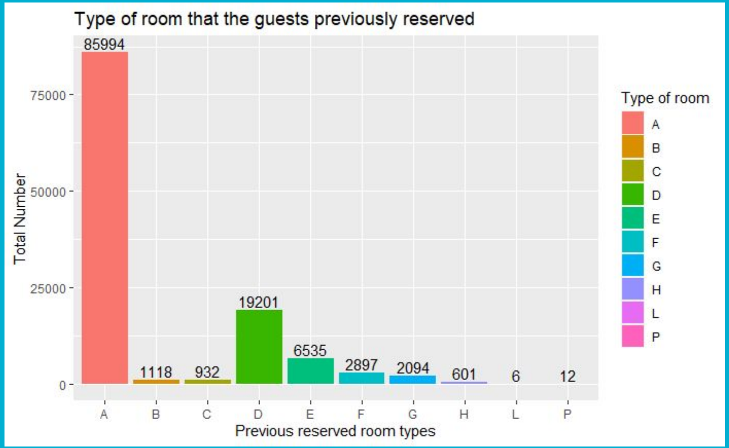

Type of Market
Total Number
Type of Marketing
Market Selection
Aviation
Complementary
Corporate
Direct
Groups
Offline TA/TO
Online TA
Undefined
237
743
5295
12606
19811
24219
56477
2
0
20000
40000

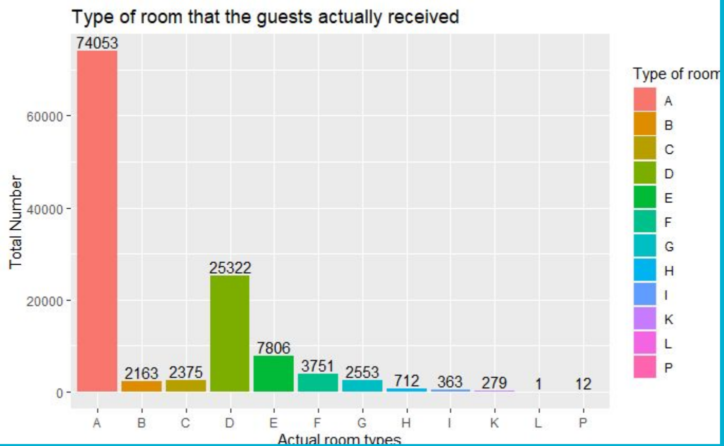

Type of Channel
Total Number
Type of Channel
Channel Distribution
Corporate
Direct
GDS
TA/TO
Undefined
6677
14645
193
97870
5
0
25000
50000
75000
100000

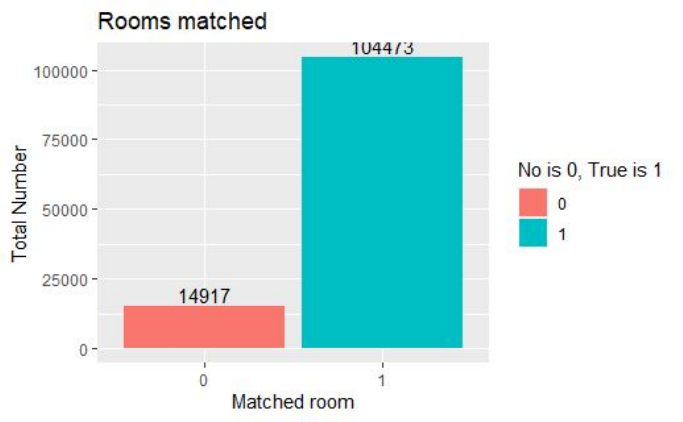

The deposit after they reserve the room
Total Number
Type of Deposit
Type of deposit
No Deposit
Non Refund
Refundable
104641
14587
162
0
25000
50000
75000
100000


Rooms are are being cancelled
Total Number
Is cancelled
0 is No, 1 is Yes
0
1
75166
44224
0
20000
40000
60000

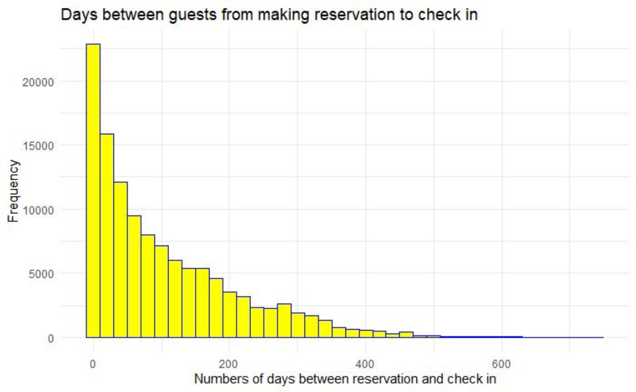
EDA for room initially reserved vs actual received

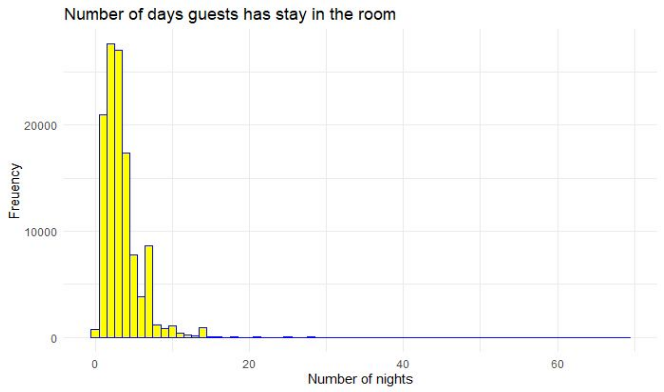


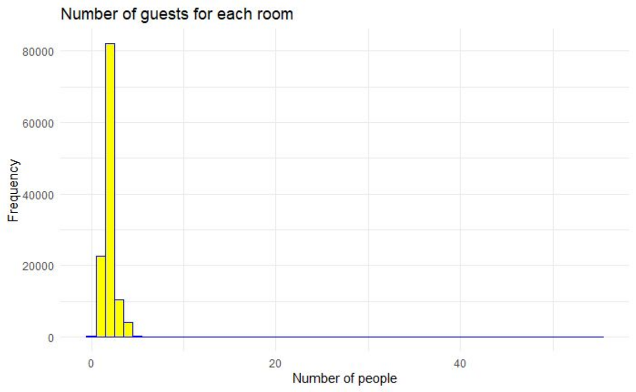


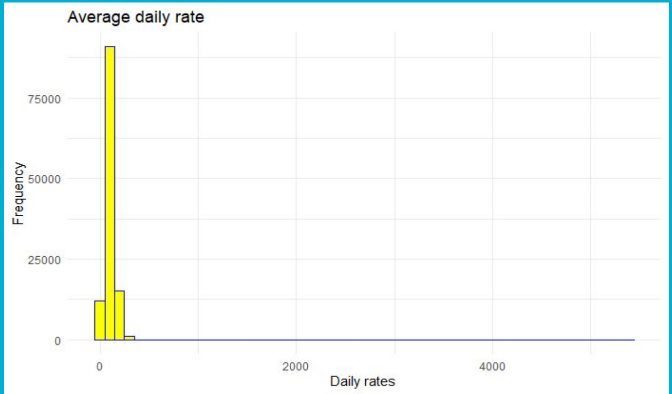


EDA for numerical variables

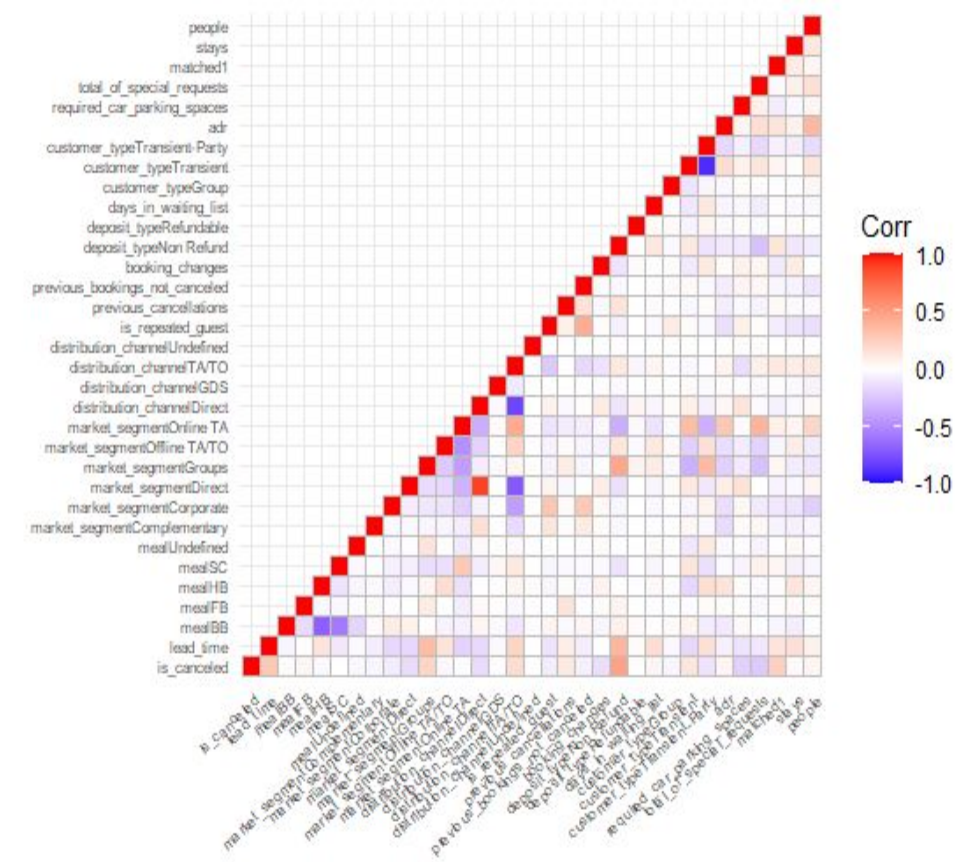








Correlation Matrix



What analysis I did use?

- I decided to use the generalized regression model analysis to determine factors that would most significantly affect people that forced to cancel their hotel room prior to check in.
- I decided to use full and reduced model to see what would be the best linear prediction that would predict the outcome of hotel cancellation.

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

VIF for 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	1433.145342
total_of_special_requests	1.184319e+00	1	1.088264
stays	1.158580e+00	1	1.076374
people	1.314950e+00	1	1.146713
matched	1.016251e+00	1	1.008093

From the VIF, I noticed that market segment, distribution channel and parking spaces required has a multicollinearity above 5, which means that it is poorly estimated of estimators, which would contain bias of determining the factors

Reduced Model

I decided to select the significant factors that selected from our full model to be used as our reduced model and remove the categories that has a VIF value that above 5.

```
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
```

Which one is better reduced or full?

```
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 +
    stays + people + matched
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
```

By typing the code anova(full_model, reduced_model), I determine that the reduced model is a better selection to determine the factors of hotel cancellations.

Final model selection:

```
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
```

```
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
```

VIF and Durbin-Watson test

I notice that the p-value is 0, which is less than alpha value of 0.05, I can conclude that I will reject the null hypothesis that using reduced model is a better predictor, conclude that the residuals in this regression model are autocorrelated.

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

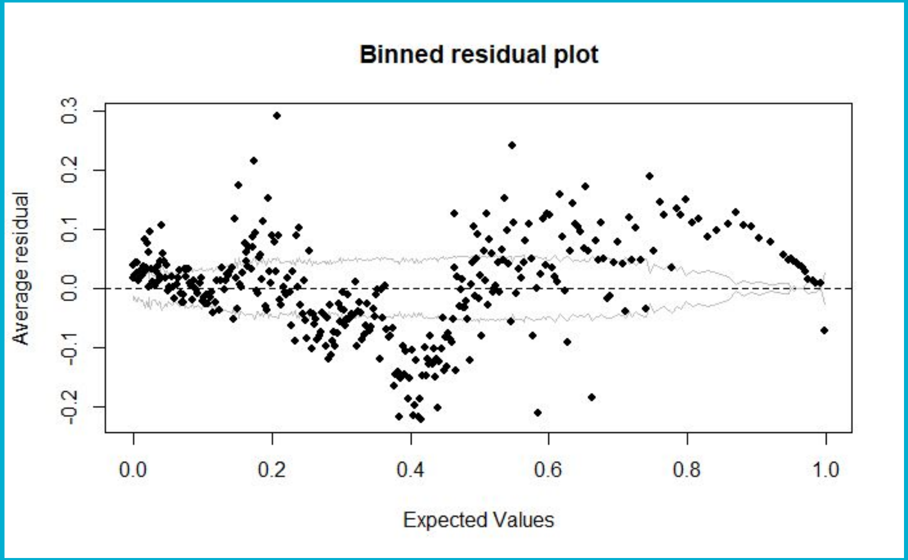
Testing AUC plot

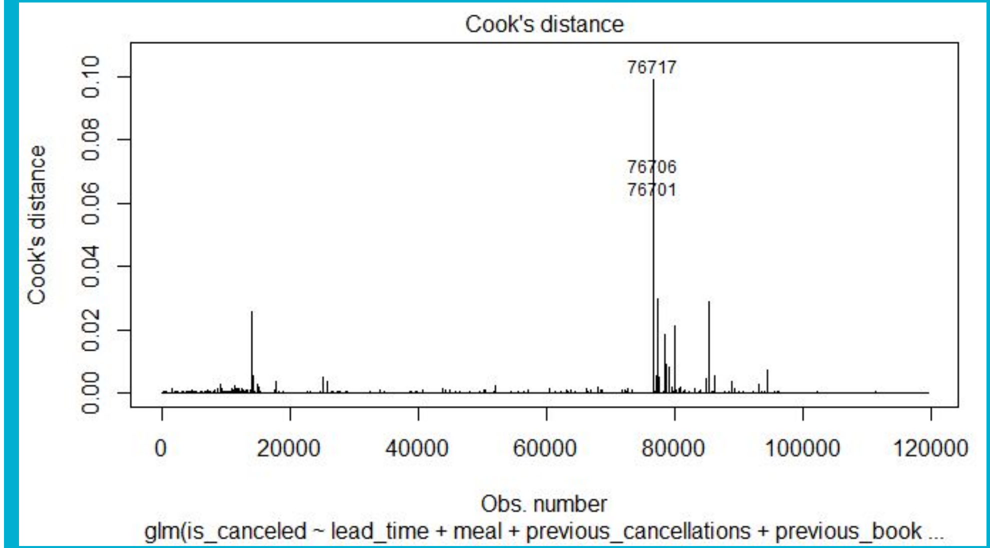
- Using the AUC function, we get 0.8182, so that means that 81.82% of my dataset has been well fitted into our dataset.

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

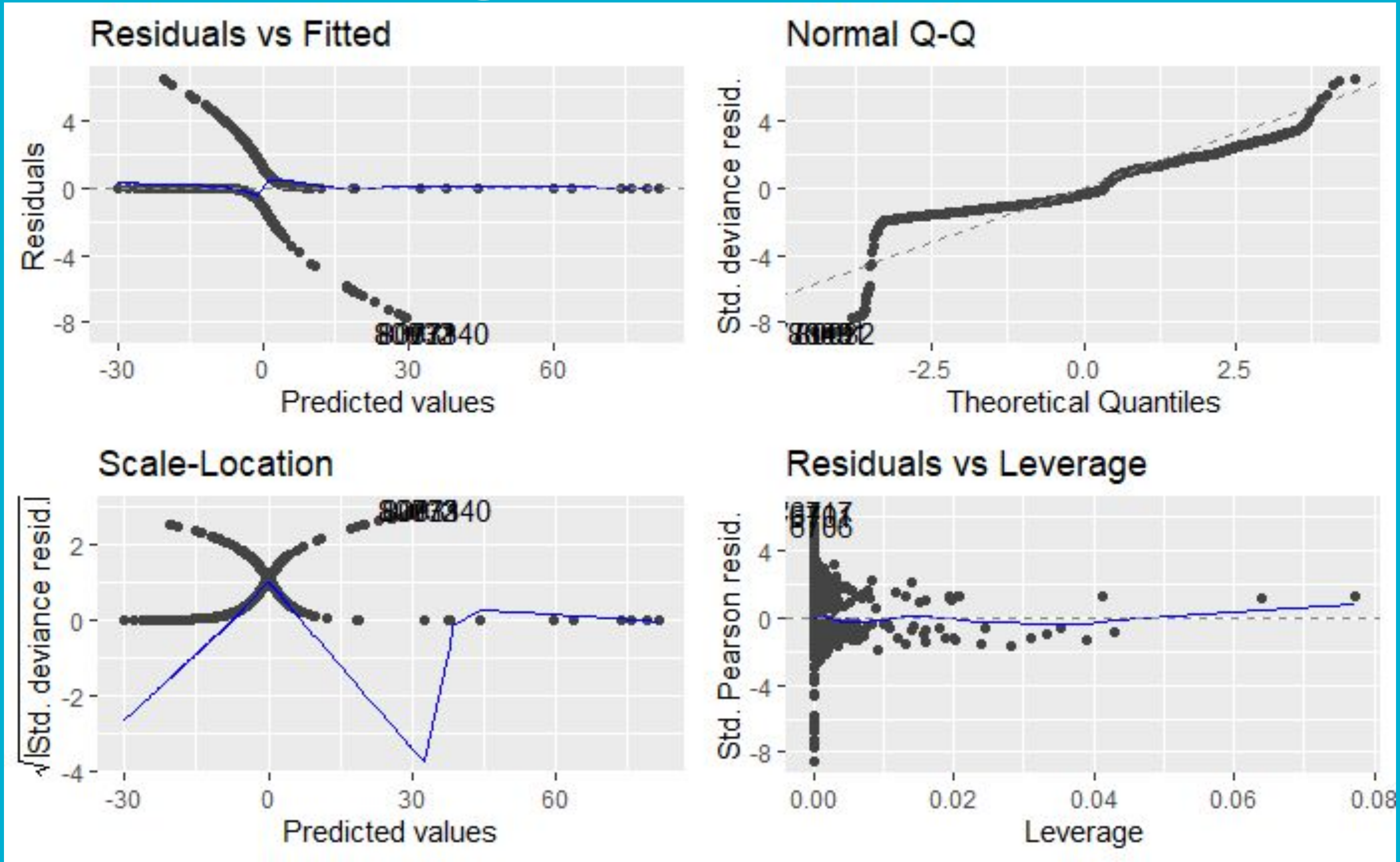
```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
Area under the curve: 0.8182
```

Residual plot and cook’s distance





Assumption testing



So what is our conclusion and limitation

- Very surprised to see there are 12 significant factors that led to their room cancelled.
- It was surprised, but at the same time, it was factual, because we will always heard a lot of reasons that they decided to cancelled their room.
- I believe that I would try to work a dataset that with even numbers of hotel type between City and Resort, and at the same time, try to keep the raw data as much as possible.
- I might find a similar hotel booking from like two to three years ago and do a comparison to this that to discover the trend of hotel cancellation for these few years.

Thank for your watching!