Regression.Rmd

```
hotel_bookings <- read.csv('C:/Users/colto/Documents/RegressionClassification/hotel_bookings.csv')
```

Source: https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

a. First we split the data set into 80/20 train/test.

```
# dividing the dataset into 80/20
n = nrow(hotel_bookings)
trainIndex = sample(1:n, size=round(0.8*n), replace=FALSE)
# assign train/test 80/20
train = hotel_bookings[trainIndex ,]
test = hotel_bookings[-trainIndex ,]
```

b. We can now explore the training data to verify it's contents.

```
head(train) # Print the first six rows
```

```
##              hotel is_canceled lead_time arrival_date_year arrival_date_month
## 32879 Resort Hotel           0       140              2017           February
## 5503  Resort Hotel           1        31              2016              April
## 80293   City Hotel           1        33              2015           December
## 8067  Resort Hotel           1        43              2016          September
## 65110   City Hotel           1        74              2017              March
## 40227   City Hotel           0        18              2015               July
##       arrival_date_week_number arrival_date_day_of_month
## 32879                        6                         6
## 5503                        18                        30
## 80293                       50                         9
## 8067                        37                         9
## 65110                       12                        24
## 40227                       29                        18
##       stays_in_weekend_nights stays_in_week_nights adults children babies meal
## 32879                       1                    1      2        0      0   BB
## 5503                        1                    1      2        0      0   HB
## 80293                       0                    2      2        0      0   BB
## 8067                        0                    2      1        0      0   BB
## 65110                       2                    2      2        0      0   BB
## 40227                       1                    1      1        0      0   BB
##       country market_segment distribution_channel is_repeated_guest
## 32879     GBR      Online TA                TA/TO                 0
## 5503      PRT         Groups                TA/TO                 0
## 80293     PRT  Offline TA/TO                TA/TO                 0
## 8067      NOR      Online TA                TA/TO                 0
```

```
## 65110     DEU      Online TA               TA/TO                0
## 40227     PRT         Groups               TA/TO                0
##       previous_cancellations previous_bookings_not_canceled reserved_room_type
## 32879                      0                              0                 E
## 5503                       0                              0                 A
## 80293                      1                              0                 A
## 8067                       0                              0                 A
## 65110                      0                              0                 E
## 40227                      0                              0                 A
##       assigned_room_type booking_changes deposit_type agent company
## 32879                  E               0   No Deposit   240    NULL
## 5503                   A               0   No Deposit   298    NULL
## 80293                  A               0   Non Refund    44    NULL
## 8067                   A               0   No Deposit   240    NULL
## 65110                  E               0   No Deposit     9    NULL
## 40227                  A               0   No Deposit     1    NULL
##       days_in_waiting_list   customer_type   adr required_car_parking_spaces
## 32879                    0       Transient  57.8                          0
## 5503                     0 Transient-Party  86.0                          0
## 80293                    0       Transient 110.0                          0
## 8067                     0       Transient 150.0                          0
## 65110                    0       Transient 117.3                          0
## 40227                    0 Transient-Party   0.0                          0
##       total_of_special_requests reservation_status reservation_status_date
## 32879                         2          Check-Out              2017-02-08
## 5503                          0           Canceled              2016-03-30
## 80293                         0           Canceled              2015-11-16
## 8067                          0           Canceled              2016-08-02
## 65110                         2           Canceled              2017-01-11
## 40227                         0          Check-Out              2015-07-20
```

```
names(train) # Get column names
```

```
##  [1] "hotel"                          "is_canceled"
##  [3] "lead_time"                      "arrival_date_year"
##  [5] "arrival_date_month"             "arrival_date_week_number"
##  [7] "arrival_date_day_of_month"      "stays_in_weekend_nights"
##  [9] "stays_in_week_nights"           "adults"
## [11] "children"                       "babies"
## [13] "meal"                           "country"
## [15] "market_segment"                 "distribution_channel"
## [17] "is_repeated_guest"              "previous_cancellations"
## [19] "previous_bookings_not_canceled" "reserved_room_type"
## [21] "assigned_room_type"             "booking_changes"
## [23] "deposit_type"                   "agent"
## [25] "company"                        "days_in_waiting_list"
## [27] "customer_type"                  "adr"
## [29] "required_car_parking_spaces"    "total_of_special_requests"
## [31] "reservation_status"             "reservation_status_date"
```

```
dim(train) # Get number of rows & columns
```

```
## [1] 95512    32
```

```
colSums(is.na(train)) # Get number of missing values
```

```
##                          hotel                    is_canceled
##                              0                              0
##                      lead_time              arrival_date_year
##                              0                              0
##             arrival_date_month       arrival_date_week_number
##                              0                              0
##      arrival_date_day_of_month        stays_in_weekend_nights
##                              0                              0
##           stays_in_week_nights                         adults
##                              0                              0
##                       children                         babies
##                              4                              0
##                           meal                        country
##                              0                              0
##                 market_segment           distribution_channel
##                              0                              0
##              is_repeated_guest         previous_cancellations
##                              0                              0
## previous_bookings_not_canceled             reserved_room_type
##                              0                              0
##             assigned_room_type                booking_changes
##                              0                              0
##                   deposit_type                          agent
##                              0                              0
##                        company           days_in_waiting_list
##                              0                              0
##                  customer_type                            adr
##                              0                              0
##    required_car_parking_spaces       total_of_special_requests
##                              0                              0
##             reservation_status         reservation_status_date
##                              0                              0
```

```
str(train) # Get structure of variables
```

```
## 'data.frame':    95512 obs. of  32 variables:
##  $ hotel                         : chr  "Resort Hotel" "Resort Hotel" "City Hotel" "Resort Hotel" ..
##  $ is_canceled                   : int  0 1 1 1 1 0 0 0 1 0 ...
##  $ lead_time                     : int  140 31 33 43 74 18 8 108 407 59 ...
##  $ arrival_date_year             : int  2017 2016 2015 2016 2017 2015 2016 2016 2017 2016 ...
##  $ arrival_date_month            : chr  "February" "April" "December" "September" ...
##  $ arrival_date_week_number      : int  6 18 50 37 12 29 45 16 18 31 ...
##  $ arrival_date_day_of_month     : int  6 30 9 9 24 18 5 14 6 24 ...
##  $ stays_in_weekend_nights       : int  1 1 0 0 2 1 0 0 2 2 ...
##  $ stays_in_week_nights          : int  1 1 2 2 2 1 1 3 1 4 ...
##  $ adults                        : int  2 2 2 1 2 1 2 2 2 2 ...
##  $ children                      : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ babies                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ meal                          : chr  "BB" "HB" "BB" "BB" ...
##  $ country                       : chr  "GBR" "PRT" "PRT" "NOR" ...
```

```
##  $ market_segment                : chr  "Online TA" "Groups" "Offline TA/TO" "Online TA" ...
##  $ distribution_channel          : chr  "TA/TO" "TA/TO" "TA/TO" "TA/TO" ...
##  $ is_repeated_guest             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ previous_cancellations        : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ previous_bookings_not_canceled: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ reserved_room_type            : chr  "E" "A" "A" "A" ...
##  $ assigned_room_type            : chr  "E" "A" "A" "A" ...
##  $ booking_changes               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ deposit_type                  : chr  "No Deposit" "No Deposit" "Non Refund" "No Deposit" ...
##  $ agent                         : chr  "240" "298" "44" "240" ...
##  $ company                       : chr  "NULL" "NULL" "NULL" "NULL" ...
##  $ days_in_waiting_list          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ customer_type                 : chr  "Transient" "Transient-Party" "Transient" "Transient" ...
##  $ adr                           : num  57.8 86 110 150 117.3 ...
##  $ required_car_parking_spaces   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ total_of_special_requests     : int  2 0 0 0 2 0 0 2 0 1 ...
##  $ reservation_status            : chr  "Check-Out" "Canceled" "Canceled" "Canceled" ...
##  $ reservation_status_date       : chr  "2017-02-08" "2016-03-30" "2015-11-16" "2016-08-02" ...
```

```
summary(train) # Basic descriptive statistics
```

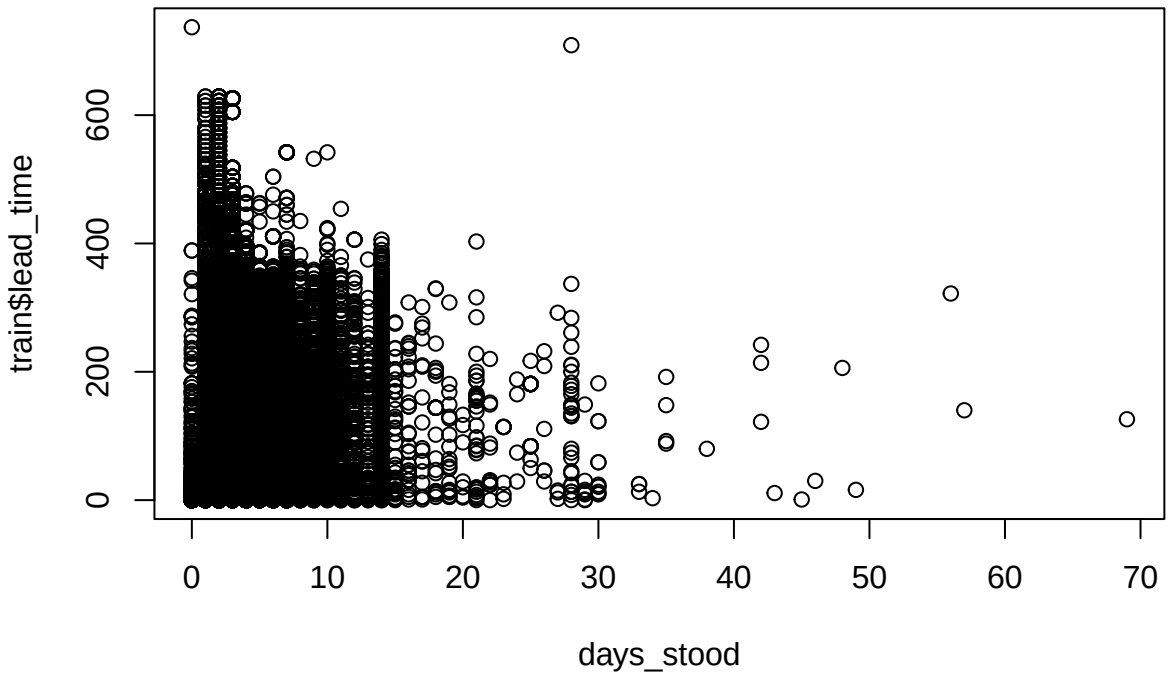
```
##     hotel            is_canceled       lead_time   arrival_date_year
##  Length:95512       Min.   :0.0000   Min.   :  0   Min.   :2015
##  Class :character   1st Qu.:0.0000   1st Qu.: 18   1st Qu.:2016
##  Mode  :character   Median :0.0000   Median : 69   Median :2016
##                     Mean   :0.3704   Mean   :104   Mean   :2016
##                     3rd Qu.:1.0000   3rd Qu.:161   3rd Qu.:2017
##                     Max.   :1.0000   Max.   :737   Max.   :2017
##
##  arrival_date_month arrival_date_week_number arrival_date_day_of_month
##  Length:95512       Min.   : 1.00            Min.   : 1.00
##  Class :character   1st Qu.:16.00            1st Qu.: 8.00
##  Mode  :character   Median :28.00            Median :16.00
##                     Mean   :27.14            Mean   :15.79
##                     3rd Qu.:38.00            3rd Qu.:23.00
##                     Max.   :53.00            Max.   :31.00
##
##  stays_in_weekend_nights stays_in_week_nights     adults
##  Min.   : 0.0000         Min.   : 0.000       Min.   : 0.000
##  1st Qu.: 0.0000         1st Qu.: 1.000       1st Qu.: 2.000
##  Median : 1.0000         Median : 2.000       Median : 2.000
##  Mean   : 0.9274         Mean   : 2.504       Mean   : 1.857
##  3rd Qu.: 2.0000         3rd Qu.: 3.000       3rd Qu.: 2.000
##  Max.   :19.0000         Max.   :50.000       Max.   :55.000
##
##     children           babies              meal             country
##  Min.   : 0.0000   Min.   : 0.000000   Length:95512       Length:95512
##  1st Qu.: 0.0000   1st Qu.: 0.000000   Class :character   Class :character
##  Median : 0.0000   Median : 0.000000   Mode  :character   Mode  :character
##  Mean   : 0.1045   Mean   : 0.008156
##  3rd Qu.: 0.0000   3rd Qu.: 0.000000
##  Max.   :10.0000   Max.   :10.000000
##  NA's   :4
##  market_segment     distribution_channel is_repeated_guest
```

```
##  Length:95512       Length:95512         Min.   :0.00000
##  Class :character   Class :character     1st Qu.:0.00000
##  Mode  :character   Mode  :character     Median :0.00000
##                                          Mean   :0.03191
##                                          3rd Qu.:0.00000
##                                          Max.   :1.00000
##
##  previous_cancellations previous_bookings_not_canceled reserved_room_type
##  Min.   : 0.00000       Min.   : 0.0000                Length:95512
##  1st Qu.: 0.00000       1st Qu.: 0.0000                Class :character
##  Median : 0.00000       Median : 0.0000                Mode  :character
##  Mean   : 0.08796       Mean   : 0.1386
##  3rd Qu.: 0.00000       3rd Qu.: 0.0000
##  Max.   :26.00000       Max.   :72.0000
##
##  assigned_room_type booking_changes deposit_type          agent
##  Length:95512       Min.   : 0.00   Length:95512       Length:95512
##  Class :character   1st Qu.: 0.00   Class :character   Class :character
##  Mode  :character   Median : 0.00   Mode  :character   Mode  :character
##                     Mean   : 0.22
##                     3rd Qu.: 0.00
##                     Max.   :21.00
##
##    company          days_in_waiting_list customer_type           adr
##  Length:95512       Min.   :  0.000      Length:95512       Min.   :   0.00
##  Class :character   1st Qu.:  0.000      Class :character   1st Qu.:  69.33
##  Mode  :character   Median :  0.000      Mode  :character   Median :  94.50
##                     Mean   :  2.272                         Mean   : 101.81
##                     3rd Qu.:  0.000                         3rd Qu.: 126.00
##                     Max.   :391.000                         Max.   :5400.00
##
##  required_car_parking_spaces total_of_special_requests reservation_status
##  Min.   :0.00000             Min.   :0.000             Length:95512
##  1st Qu.:0.00000             1st Qu.:0.000             Class :character
##  Median :0.00000             Median :0.000             Mode  :character
##  Mean   :0.06229             Mean   :0.572
##  3rd Qu.:0.00000             3rd Qu.:1.000
##  Max.   :8.00000             Max.   :5.000
##
##  reservation_status_date
##  Length:95512
##  Class :character
##  Mode  :character
##
##
##
##
```

c. We can also create informative graphs using the data.

Is there some sort of relationship between the number of days stood and the lead time?

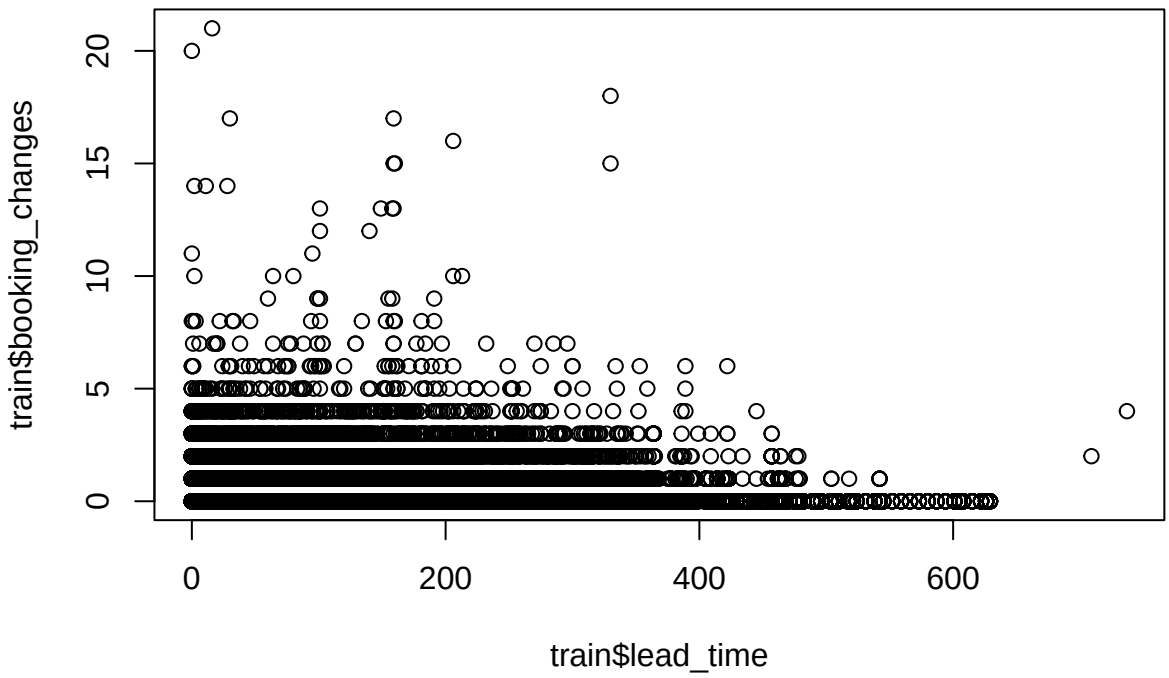
```r
# Add stays_in_week_nights and stays_in_weekend_nights
days_stood = train$stays_in_week_nights+train$stays_in_weekend_nights
# Plot days_stood v lead_time
plot(days_stood, train$lead_time)
```



Right now, it's difficult to tell if there is due to the density of the points. This isn't suprising, since we can expect the vast majority of points to be distributed towards x=0 as extremely long stays at a hotel are unusual and uncommon. However, we can observe that those who give 400+ days of lead time tend to not stay for very long.(Less than 10 days) We can also observe that those who stay for very long (10 or more days) tend to give a year or less of lead time. My hypothesis is that there is an inverse correlation between the number of days stood and the lead time. We will build a linear regression model to test this hypothesis later.

Is there some sort of relationship between the lead time and the number of booking changes?

```r
# Add stays_in_week_nights and stays_in_weekend_nights
stays = train$stays_in_week_nights+train$stays_in_weekend_nights
# Plot lead_time v booking_changes
plot(train$lead_time, train$booking_changes)
```



Right now, it's difficult to tell if there is due to the density of the points, but the outliers do provide a hint that there may be an inverse correlation between the lead time and the number of booking changes. We won't analyze this relationship any further as it is not required, but my hypothesis is that there is no correlation between these two attributes.

d. We'll make a linear regression model for the relationship between the duration of the hotel stay and the lead time.

```
# Add stays_in_week_nights and stays_in_weekend_nights
stays = train$stays_in_week_nights+train$stays_in_weekend_nights
# lm between lead_time and stays
results1 <- lm(lead_time ~ stays, data=train)
# Display summary of the results
summary(results1)
```

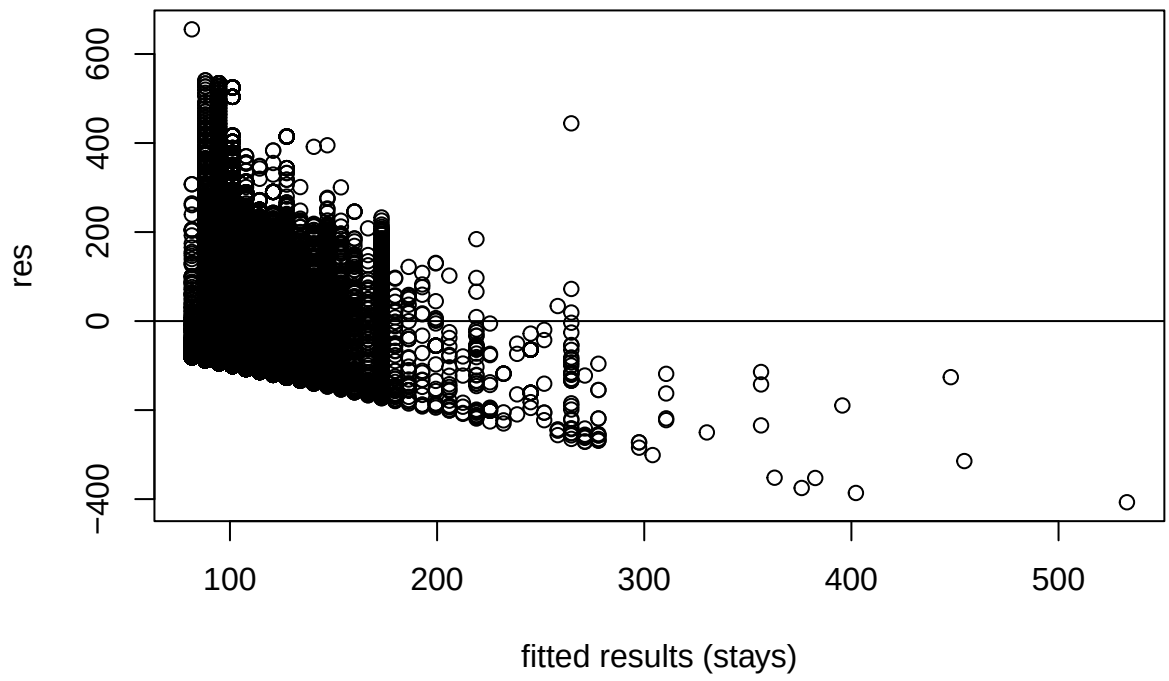
```
##
## Call:
## lm(formula = lead_time ~ stays, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -407.03  -81.68  -35.68   50.78  655.41
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  81.5948     0.5718  142.71   <2e-16 ***
## stays         6.5425     0.1336   48.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 105.5 on 95510 degrees of freedom
## Multiple R-squared:  0.02448,    Adjusted R-squared:  0.02447
## F-statistic:  2396 on 1 and 95510 DF,  p-value: < 2.2e-16
```

First of all, our residuals are interesting since the median is far from 0. This is likely due to the high values of the rest of the residuals, as well as the high residual standard error. The rest of our residuals look okay since they somewhat reflect each other, we can again attribute the imperfections to our high residual standard error. Since both P-values are considerably lower than 0.05, we can say that the amount of days stood is a significant factor to the lead time. According to the coefficient estimate, there is a positive correlation between the number of days stood and the amount of lead time (Estimate Std. > 0), disproving my hypothesis from part C. The value of R squared suggests that 2.5% of the variance in lead time can be attributed to the length of the hotel stay.

e. Now we'll plot the residuals.

```r
# Get list of residuals
res <- resid(results1)
# Produce residual vs. fitted plot
plot(fitted(results1), res, xlab="fitted results (stays)")
# Add horizontal line at 0
abline(0,0)
```



According to the graph, the spread of the residuals tends to be of lower fitted values, approaching y=0. Residual outliers tend to be of lower value tend and of higher fitted value. The downward trend suggests that there may be a better way to model the relationship, though we cannot say for certain due to the randomness of the model around the residual line and the density of the points is difficult to perceive.

f. Now we'll do a linear regression model with multiple predictors.

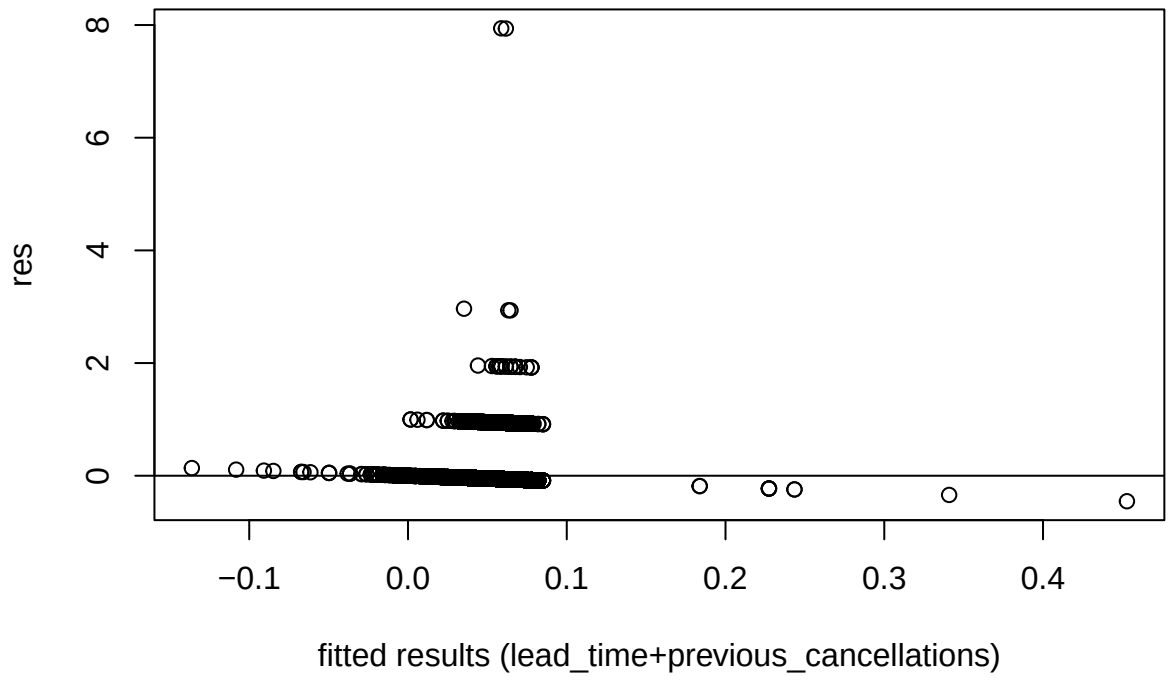
Is the required number of parking spaces related to the number of adults and the duration of the hotel stay?

```
# Add stays_in_week_nights and stays_in_weekend_nights
stays = train$stays_in_week_nights+train$stays_in_weekend_nights
# lm between lead_time, stays, and
results2 <- lm(required_car_parking_spaces ~ adults+stays, data=train)
# Display summary of the results
summary(results2)
```

```
##
## Call:
## lm(formula = required_car_parking_spaces ~ adults + stays, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4529 -0.0675 -0.0632 -0.0573  7.9413
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0588435  0.0028072  20.962  < 2e-16 ***
## adults       0.0072705  0.0013928   5.220 1.79e-07 ***
## stays       -0.0029314  0.0003125  -9.381  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2454 on 95509 degrees of freedom
## Multiple R-squared:  0.001109,   Adjusted R-squared:  0.001088
## F-statistic: 53.03 on 2 and 95509 DF,  p-value: < 2.2e-16
```

g. Now we'll plot the residuals.

```
# Get list of residuals
res <- resid(results2)
# Produce residual vs. fitted plot
plot(fitted(results2), res, xlab="fitted results (lead_time+previous_cancellations)")
# Add horizontal line at 0
abline(0,0)
```



g. We'll improve upon the linear regression model from part d. This time we'll use arrival_date_week_number and is_repeated_guest as extra predictors. The goal is to improve the previous R squared of 2.5%.

```
# Add stays_in_week_nights and stays_in_weekend_nights
stays = train$stays_in_week_nights+train$stays_in_weekend_nights
# lm between lead_time and stays
results3 <- lm(lead_time ~ stays+arrival_date_week_number+is_repeated_guest, data=train)
# Display summary of the results
summary(results3)
```

```
##
## Call:
## lm(formula = lead_time ~ stays + arrival_date_week_number + is_repeated_guest,
##     data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -371.67  -76.81  -30.63   50.72  651.48
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                59.51774    0.88019   67.62   <2e-16 ***
## stays                       5.95924    0.13259   44.94   <2e-16 ***
## arrival_date_week_number    0.96299    0.02477   38.88   <2e-16 ***
## is_repeated_guest         -64.59738    1.92817  -33.50   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 104.1 on 95508 degrees of freedom
## Multiple R-squared:  0.05142,    Adjusted R-squared:  0.05139
## F-statistic:  1726 on 3 and 95508 DF,  p-value: < 2.2e-16
```

h. The new R squared is 5.1%, so this is an improvement. Though, it is difficult to say for certain since the model we're trying to improve had a very low R squared (2.5%), which suggests that the two attributes are hardly related to each other. However, this is to be expected because the variables that influence the lead time are far too numerous and vast majority of them are not encompassed by data set. Therefore, this new model is an improvement.

i. We need to remove the non-numeric columns from our data set, then we can create the correlation matrix. (metrics correlation) In addition, we will print the MSE of each model.

```
test_copy = subset(test, select=-c(hotel,arrival_date_month,meal,country,market_segment,distribution_cha
test_copy.cor = cor(test_copy) # cor matrix
summary(test_copy.cor) # print matrix
```

```
##   is_canceled          lead_time        arrival_date_year
##  Min.   :-0.2252622   Min.   :-0.11979   Min.   :-0.548518
##  1st Qu.:-0.0561124   1st Qu.:-0.05642   1st Qu.:-0.003691
##  Median :-0.0008311   Median : 0.02685   Median : 0.026989
##  Mean   : 0.0463724   Mean   : 0.08859   Mean   : 0.047393
##  3rd Qu.: 0.0452733   3rd Qu.: 0.11907   3rd Qu.: 0.039454
##  Max.   : 1.0000000   Max.   : 1.00000   Max.   : 1.000000
##  arrival_date_week_number arrival_date_day_of_month stays_in_weekend_nights
##  Min.   :-0.548518        Min.   :-0.0331426        Min.   :-0.08060
##  1st Qu.: 0.002237        1st Qu.:-0.0086509        1st Qu.:-0.01778
##  Median : 0.013869        Median : 0.0001856        Median : 0.02543
##  Mean   : 0.045237        Mean   : 0.0575684        Mean   : 0.09730
##  3rd Qu.: 0.030783        3rd Qu.: 0.0121363        3rd Qu.: 0.07679
##  Max.   : 1.000000        Max.   : 1.0000000        Max.   : 1.00000
##  stays_in_week_nights     adults            children            babies
##  Min.   :-0.08964     Min.   :-0.14084   Min.   :-0.03562   Min.   :-0.039503
##  1st Qu.:-0.01063     1st Qu.:-0.00568   1st Qu.:-0.01677   1st Qu.:-0.007885
##  Median : 0.03087     Median : 0.02633   Median : 0.02734   Median : 0.014346
##  Mean   : 0.10749     Mean   : 0.08382   Mean   : 0.08804   Mean   : 0.069806
##  3rd Qu.: 0.07908     3rd Qu.: 0.09038   3rd Qu.: 0.04506   3rd Qu.: 0.028392
##  Max.   : 1.00000     Max.   : 1.00000   Max.   : 1.00000   Max.   : 1.000000
##  is_repeated_guest  previous_cancellations previous_bookings_not_canceled
##  Min.   :-0.14084   Min.   :-0.11986       Min.   :-0.10207
##  1st Qu.:-0.08353   1st Qu.:-0.02505       1st Qu.:-0.04611
##  Median :-0.01619   Median :-0.01045       Median :-0.00761
##  Mean   : 0.04979   Mean   : 0.06132       Mean   : 0.06960
##  3rd Qu.: 0.01689   3rd Qu.: 0.06895       3rd Qu.: 0.04642
##  Max.   : 1.00000   Max.   : 1.00000       Max.   : 1.00000
##  booking_changes     days_in_waiting_list      adr
##  Min.   :-0.140530   Min.   :-0.083673    Min.   :-0.12808
##  1st Qu.: 0.008785   1st Qu.:-0.034310    1st Qu.:-0.02949
##  Median : 0.020370   Median :-0.009654    Median : 0.05329
##  Mean   : 0.070282   Mean   : 0.050247    Mean   : 0.11037
```

```
##  3rd Qu.: 0.052867   3rd Qu.: 0.017268    3rd Qu.: 0.16290
##  Max.   : 1.000000   Max.   : 1.000000    Max.   : 1.00000
##  required_car_parking_spaces total_of_special_requests
##  Min.   :-0.19909            Min.   :-0.22526
##  1st Qu.:-0.01498            1st Qu.: 0.01342
##  Median : 0.01425            Median : 0.06464
##  Mean   : 0.06120            Mean   : 0.08757
##  3rd Qu.: 0.06832            3rd Qu.: 0.10257
##  Max.   : 1.00000            Max.   : 1.00000
```

```
mean(results1$residuals^2) # MSE of model 1
```

```
## [1] 11138.57
```

```
mean(results2$residuals^2) # MSE of model 2
```

```
## [1] 0.06020269
```

```
mean(results3$residuals^2) # MSE of model 3
```

```
## [1] 10830.98
```

Unsurprisingly, the MSE of model 2 is different from model 1 and model 3 because we're testing different targets. The MSE of model 2 is healthy, as a value close to 0 suggests that is a good predictor.

The MSE of 1 and 3 suggests that our predictors are poor predictors for the target, but the MSE does suggest that model 3 is an improvement from model 1.

Student Note: If the prof wanted all these parts to use same target, she could have specified it in the doc. It took until #3 (Classification.Rmd) for me to consider that the prof might have intended for each part to be tested for the same target. In such case, could you be easy on my score where such a factor matters? I sunk hours into this project before I realized that using the same target over each part might have been the prof's intention.