Hotel Reservations and Cancellations

Reese Putnam

2025-01-30

Over the past couple years hotels have experinced a higher volume of hotel cancellations and no-show reservations resulting in a loss of revenue. Many reservations are cancelled last minute due to free cancellations or offered to them at a lower cost by these hotels. We will evalute the following data set to see if there are outlining reasons that can lead customers to cancelling a reservation.

The dataset that we will be looking at is the Hotel Reservations data set showing various data on the type of booking along with how many guests to what meal type they have reserved. We will see what factors that could be affecting their decision to either go through with their booking or end up cancelling it. The data shows bookings from 2017 and 2018 made publicly avaiable through Kaggle.

```
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
install.packages("lubridate")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
install.packages("janitor")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
install.packages("skimr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
install.packages("here")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
install.packages("readr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(ggplot2)
library(lubridate)
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(skimr)
library(here)
```

```
## here() starts at /cloud/project
```

```
library(readr)
```

```
data <- read_csv("Hotel Reservations/hotel_reservations.csv")
```

```
## Rows: 36275 Columns: 19
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## chr  (5): Booking_ID, type_of_meal_plan, room_type_reserved, market_segment_...
## dbl (14): no_of_adults, no_of_children, no_of_weekend_nights, no_of_week_nig...
##
## 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.
```

```
str(data)
```

```
## spc_tbl_ [36,275 × 19] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Booking_ID                          : chr [1:36275] "INN00001" "INN00002" "INN00003" "INN0
0004" ...
##  $ no_of_adults                        : num [1:36275] 2 2 1 2 2 2 2 2 3 2 ...
##  $ no_of_children                      : num [1:36275] 0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_weekend_nights                : num [1:36275] 1 2 2 0 1 0 1 1 0 0 ...
##  $ no_of_week_nights                   : num [1:36275] 2 3 1 2 1 2 3 3 4 5 ...
##  $ type_of_meal_plan                   : chr [1:36275] "Meal Plan 1" "Not Selected" "Meal Pla
n 1" "Meal Plan 1" ...
##  $ required_car_parking_space          : num [1:36275] 0 0 0 0 0 0 0 0 0 0 ...
##  $ room_type_reserved                  : chr [1:36275] "Room_Type 1" "Room_Type 1" "Room_Type
1" "Room_Type 1" ...
##  $ lead_time                           : num [1:36275] 224 5 1 211 48 346 34 83 121 44 ...
##  $ arrival_year                        : num [1:36275] 2017 2018 2018 2018 2018 ...
##  $ arrival_month                       : num [1:36275] 10 11 2 5 4 9 10 12 7 10 ...
##  $ arrival_date                        : num [1:36275] 2 6 28 20 11 13 15 26 6 18 ...
##  $ market_segment_type                 : chr [1:36275] "Offline" "Online" "Online" "Online"
...
##  $ repeated_guest                      : num [1:36275] 0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_previous_cancellations        : num [1:36275] 0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_previous_bookings_not_canceled: num [1:36275] 0 0 0 0 0 0 0 0 0 0 ...
##  $ avg_price_per_room                  : num [1:36275] 65 106.7 60 100 94.5 ...
##  $ no_of_special_requests              : num [1:36275] 0 1 0 0 0 1 1 1 1 3 ...
##  $ booking_status                      : chr [1:36275] "Not_Canceled" "Not_Canceled" "Cancele
d" "Canceled" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Booking_ID = col_character(),
##   ..   no_of_adults = col_double(),
##   ..   no_of_children = col_double(),
##   ..   no_of_weekend_nights = col_double(),
##   ..   no_of_week_nights = col_double(),
##   ..   type_of_meal_plan = col_character(),
##   ..   required_car_parking_space = col_double(),
##   ..   room_type_reserved = col_character(),
##   ..   lead_time = col_double(),
##   ..   arrival_year = col_double(),
##   ..   arrival_month = col_double(),
##   ..   arrival_date = col_double(),
##   ..   market_segment_type = col_character(),
##   ..   repeated_guest = col_double(),
##   ..   no_of_previous_cancellations = col_double(),
##   ..   no_of_previous_bookings_not_canceled = col_double(),
##   ..   avg_price_per_room = col_double(),
##   ..   no_of_special_requests = col_double(),
##   ..   booking_status = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
summary(data)
```

```
##   Booking_ID         no_of_adults   no_of_children    no_of_weekend_nights
##  Length:36275       Min.   :0.000   Min.   : 0.0000   Min.   :0.0000
##  Class :character   1st Qu.:2.000   1st Qu.: 0.0000   1st Qu.:0.0000
##  Mode  :character   Median :2.000   Median : 0.0000   Median :1.0000
##                     Mean   :1.845   Mean   : 0.1053   Mean   :0.8107
##                     3rd Qu.:2.000   3rd Qu.: 0.0000   3rd Qu.:2.0000
##                     Max.   :4.000   Max.   :10.0000   Max.   :7.0000
##  no_of_week_nights type_of_meal_plan  required_car_parking_space
##  Min.   : 0.000    Length:36275       Min.   :0.00000
##  1st Qu.: 1.000    Class :character   1st Qu.:0.00000
##  Median : 2.000    Mode  :character   Median :0.00000
##  Mean   : 2.204                       Mean   :0.03099
##  3rd Qu.: 3.000                       3rd Qu.:0.00000
##  Max.   :17.000                       Max.   :1.00000
##  room_type_reserved   lead_time       arrival_year  arrival_month
##  Length:36275       Min.   :  0.00   Min.   :2017   Min.   : 1.000
##  Class :character   1st Qu.: 17.00   1st Qu.:2018   1st Qu.: 5.000
##  Mode  :character   Median : 57.00   Median :2018   Median : 8.000
##                     Mean   : 85.23   Mean   :2018   Mean   : 7.424
##                     3rd Qu.:126.00   3rd Qu.:2018   3rd Qu.:10.000
##                     Max.   :443.00   Max.   :2018   Max.   :12.000
##   arrival_date  market_segment_type repeated_guest
##  Min.   : 1.0   Length:36275        Min.   :0.00000
##  1st Qu.: 8.0   Class :character    1st Qu.:0.00000
##  Median :16.0   Mode  :character    Median :0.00000
##  Mean   :15.6                       Mean   :0.02564
##  3rd Qu.:23.0                       3rd Qu.:0.00000
##  Max.   :31.0                       Max.   :1.00000
##  no_of_previous_cancellations no_of_previous_bookings_not_canceled
##  Min.   : 0.00000             Min.   : 0.0000
##  1st Qu.: 0.00000             1st Qu.: 0.0000
##  Median : 0.00000             Median : 0.0000
##  Mean   : 0.02335             Mean   : 0.1534
##  3rd Qu.: 0.00000             3rd Qu.: 0.0000
##  Max.   :13.00000             Max.   :58.0000
##  avg_price_per_room no_of_special_requests booking_status
##  Min.   :  0.00     Min.   :0.0000         Length:36275
##  1st Qu.: 80.30     1st Qu.:0.0000         Class :character
##  Median : 99.45     Median :0.0000         Mode  :character
##  Mean   :103.42     Mean   :0.6197
##  3rd Qu.:120.00     3rd Qu.:1.0000
##  Max.   :540.00     Max.   :5.0000
```

Now that the packages needed, data needed, and we have seen the perimeters of the data set we can go ahead with cleaning the data before we begin with the visualizations.

```
colSums(is.na(data))
```

```
##                          Booking_ID                        no_of_adults
##                                   0                                   0
##                       no_of_children                no_of_weekend_nights
##                                   0                                   0
##                     no_of_week_nights                   type_of_meal_plan
##                                   0                                   0
##           required_car_parking_space                   room_type_reserved
##                                   0                                   0
##                            lead_time                        arrival_year
##                                   0                                   0
##                         arrival_month                        arrival_date
##                                   0                                   0
##                  market_segment_type                      repeated_guest
##                                   0                                   0
##         no_of_previous_cancellations no_of_previous_bookings_not_canceled
##                                   0                                   0
##                   avg_price_per_room               no_of_special_requests
##                                   0                                   0
##                       booking_status
##                                   0
```

```
data <- na.omit(data)
```

We have checked and removed any rows with missing values

```
duplicated_rows <- duplicated(data)
```

```
data <- data %>%
    rename(arrival_day_of_month= arrival_date )
```

```
data <- data%>%
    mutate(arrival_date = as.Date(paste(arrival_year, arrival_month, arrival_day_of_month, sep
="-")))
```

```
data <- data %>%
  mutate(day_of_week = format(arrival_date, "%a"))
```

Next we need to add the arrival month name to make the visualizing more visible.

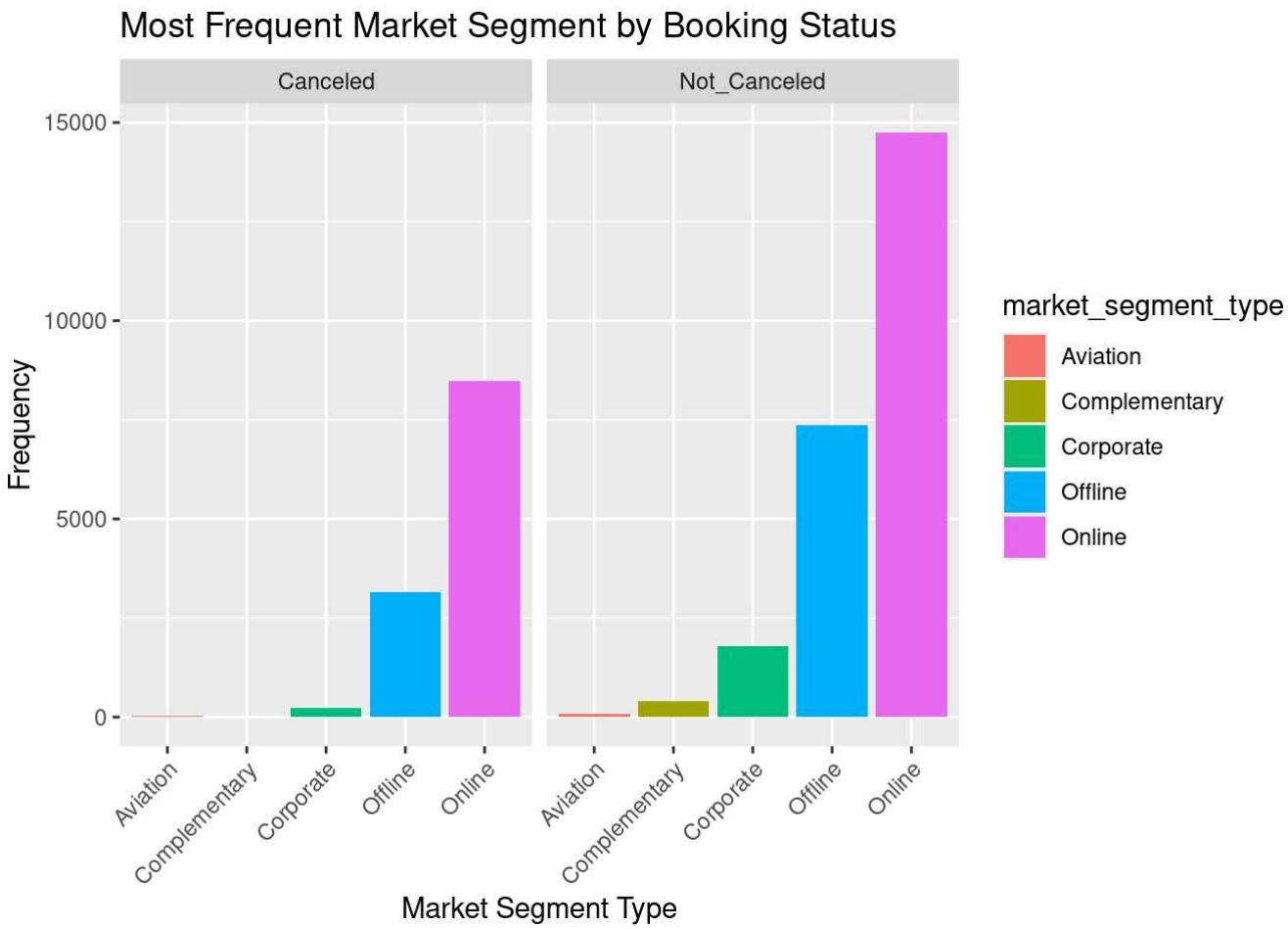
```
data<- data %>%
    mutate(arrival_month_name = month.name[arrival_month])
```

Now that the data has been cleaned and the dates have been formatted for easier visibility we can begin visualizing the data to see what factors contribute to guests cancelling their hotel reservations.

```
ggplot(data, aes(x= booking_status, y= lead_time, fill= booking_status))+
  geom_bar(stat="identity")+
  labs(title = "Lead Time by Booking Status", x= "Booking Status", y= "Lead Time")
```

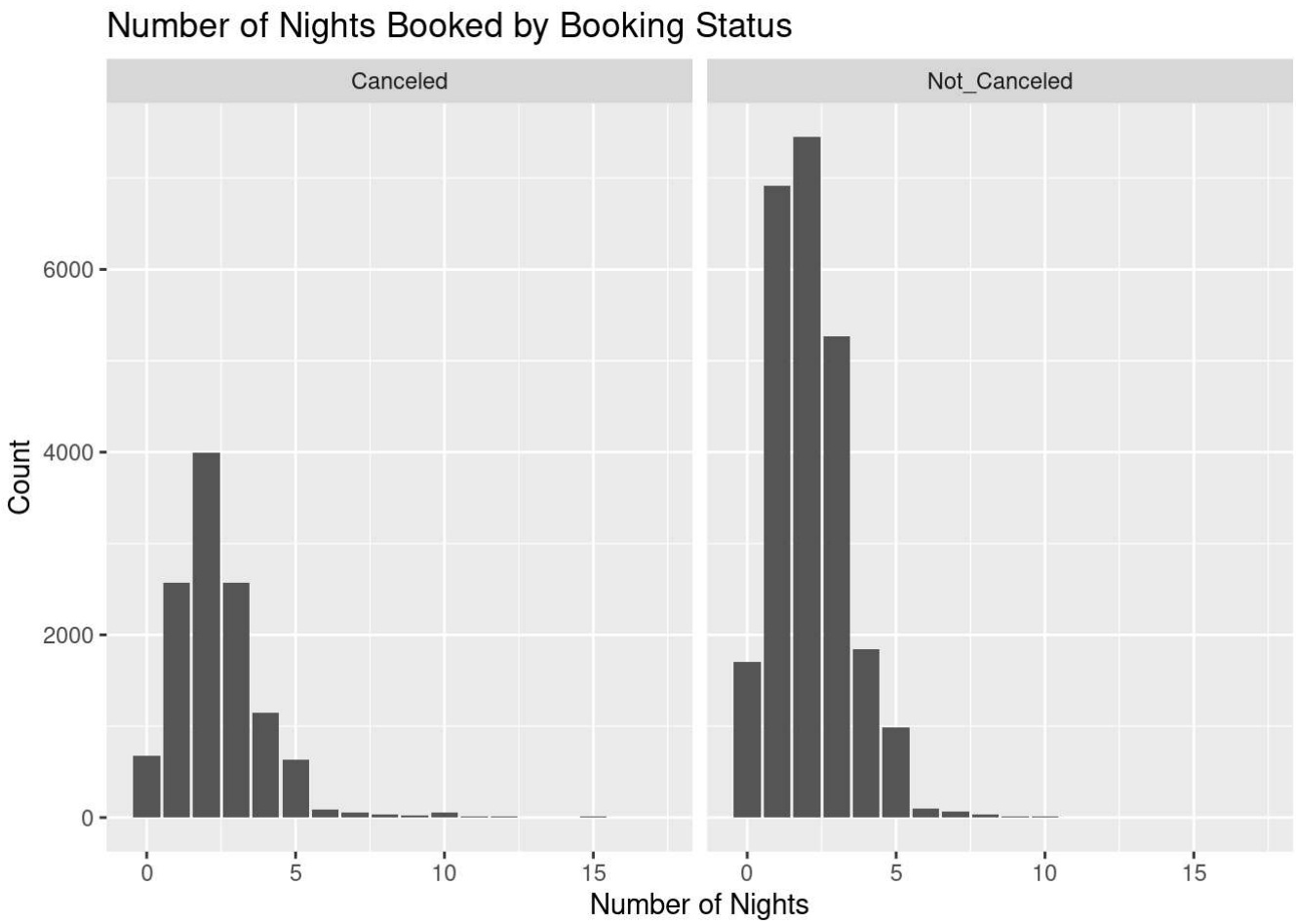
From the chart above we can see that bookings made with a larger lead time have more of a chance of being cancelled than reservations that are booked closer to the date of their stay.

```
ggplot(data , aes(x = market_segment_type, fill = market_segment_type))+
    geom_bar()+
    facet_wrap(~booking_status)+
    labs(title = "Most Frequent Market Segment by Booking Status",
         x = "Market Segment Type",
         y = "Frequency")+
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



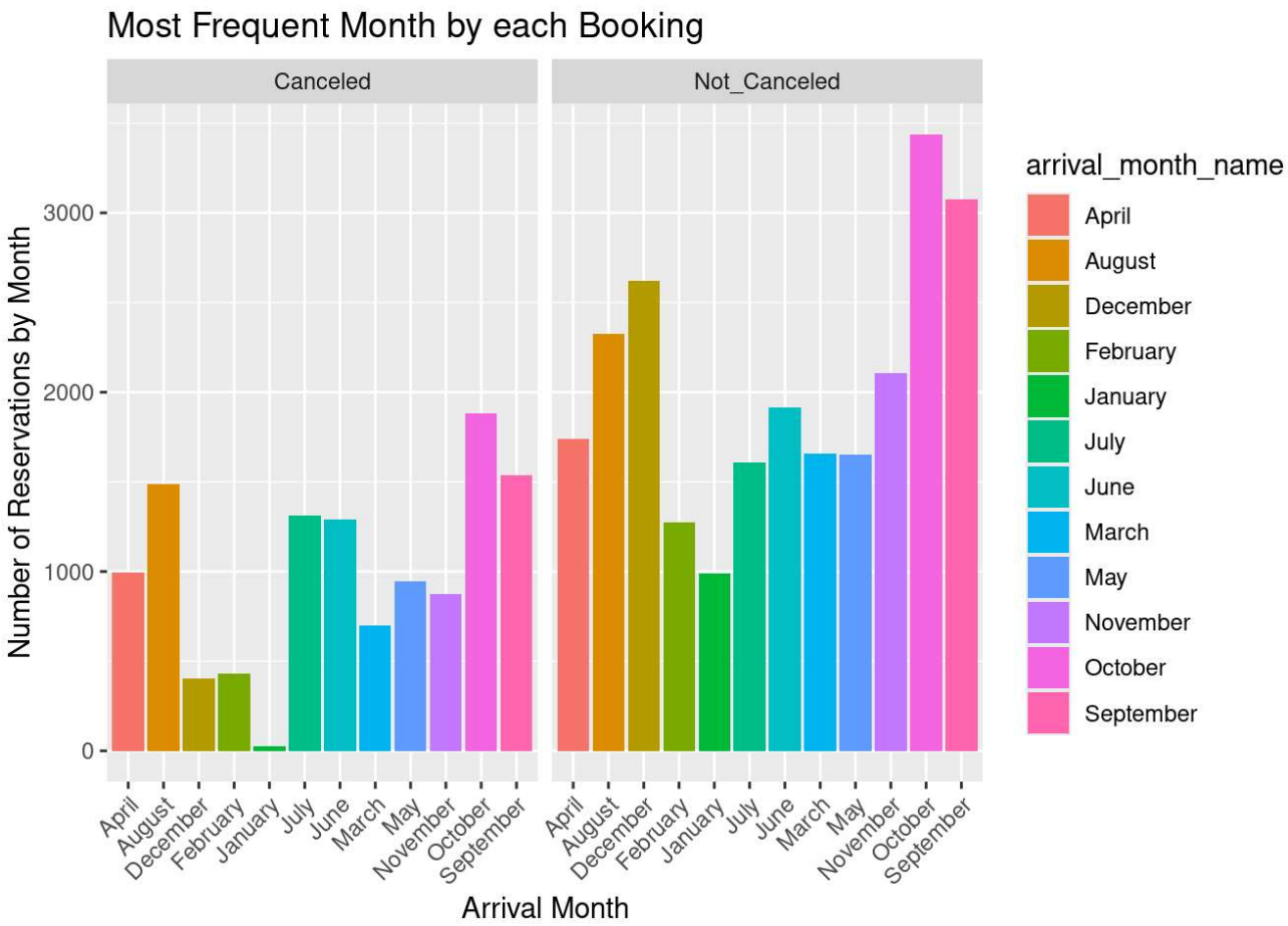
Based off the market segment chart online bookings are more likely to be cancelled than bookings through different platforms but it doesnt show a significant difference since most online bookings aren't cancelled.

```
ggplot(data, aes(x= no_of_week_nights))+
  geom_bar()+
  labs(title= "Number of Nights Booked by Booking Status", x="Number of Nights", y= "Count")+
  facet_wrap(~booking_status)
```



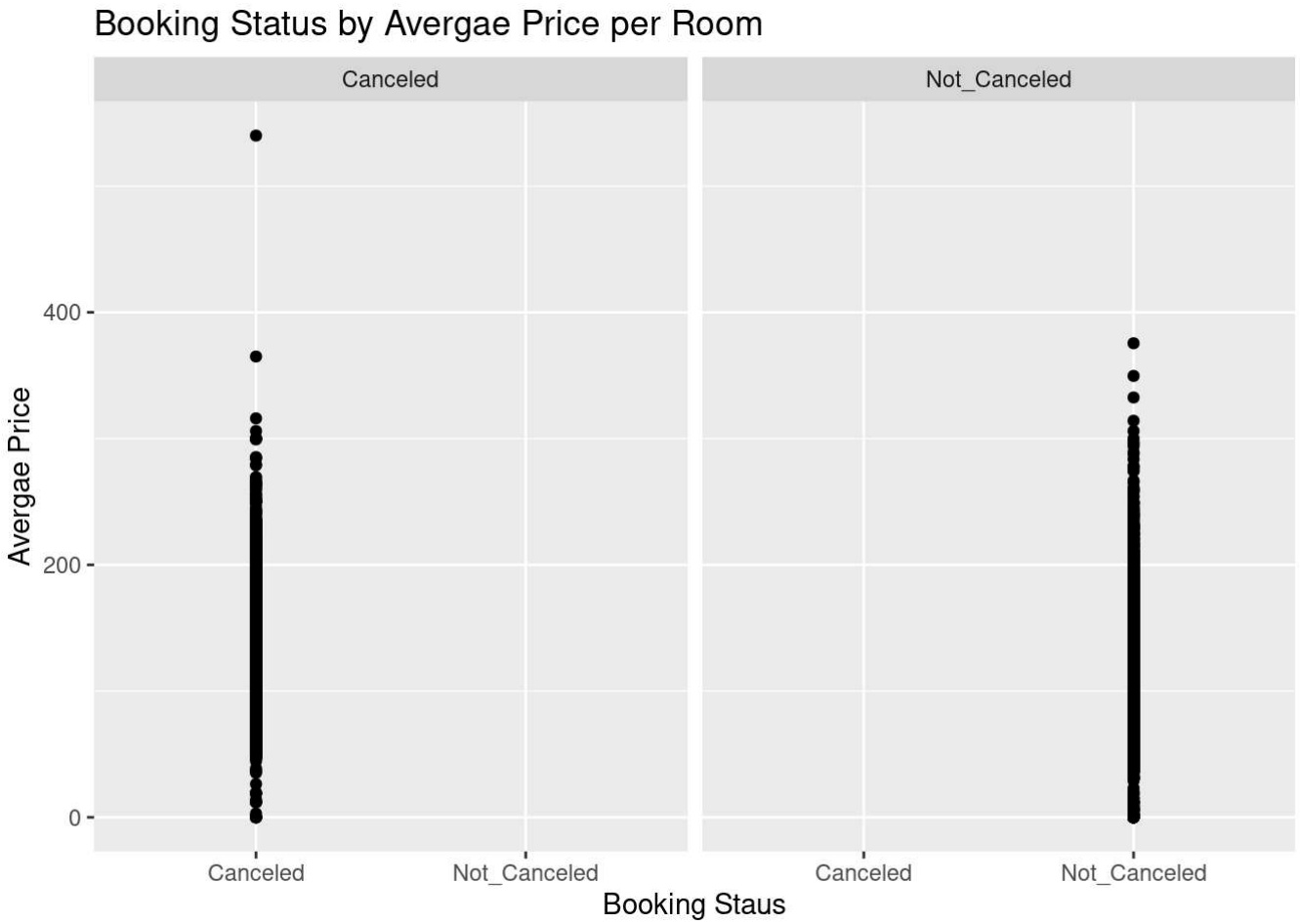
There is no trend shown above that the number of nights would effect the booking status.

```
ggplot(data, aes(x= arrival_month_name, fill=arrival_month_name))+
  geom_bar(stat="Count")+
  labs(title= "Most Frequent Month by each Booking", x= "Arrival Month", y= "Number of Reservati
ons by Month")+
  facet_wrap(~booking_status)+
  theme(axis.text.x = element_text (angle = 45, hjust =1 ))
```



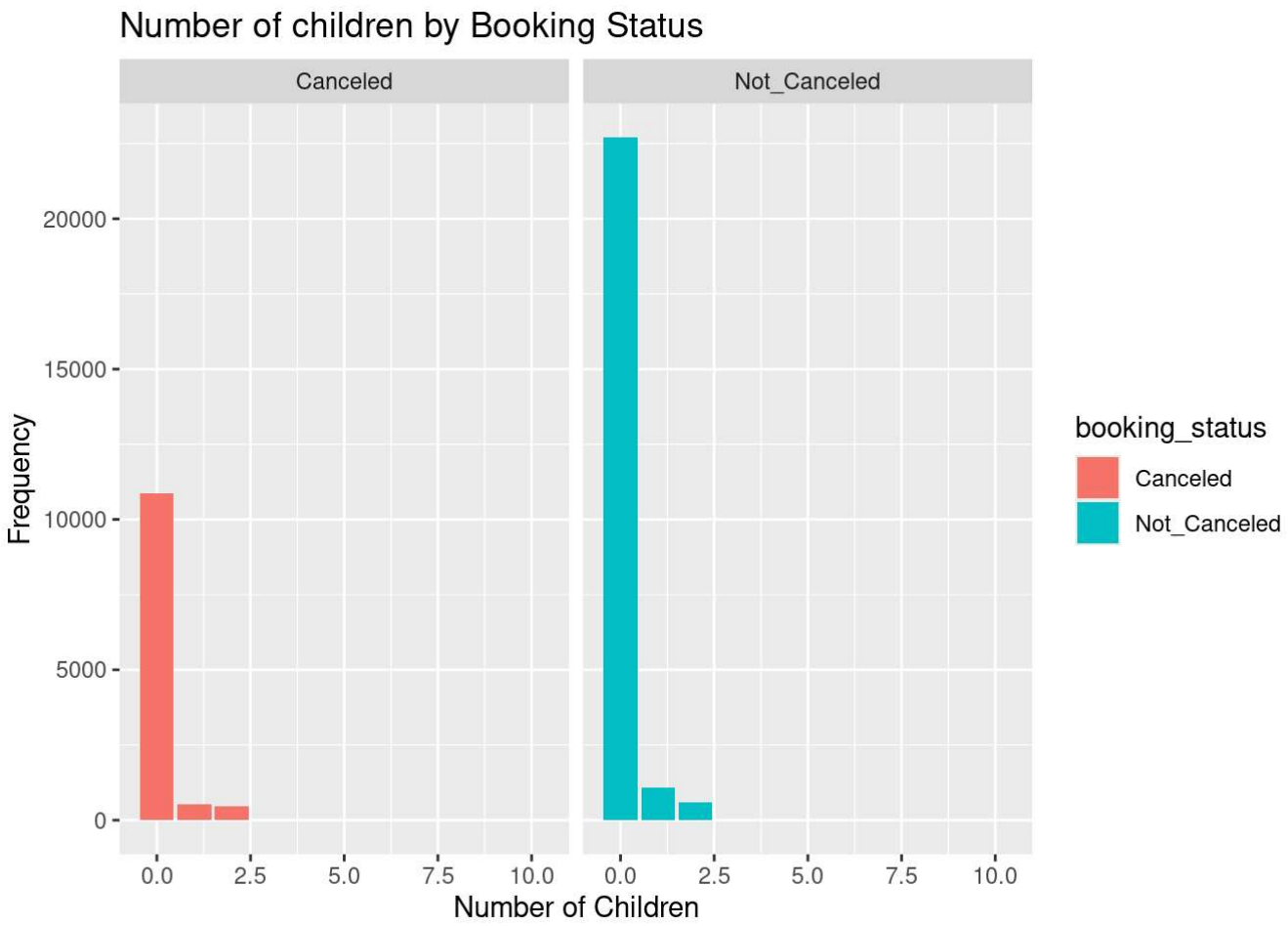
There is shown to be an increase of cancellations in the months of August, October, and September but there is a much larger number of bookings that are not cancelled in the same months. So there shows to be a trend in busier months to have a reasonable amount of cancellations for the large number of overall bookings.

```
ggplot(data, aes(x=booking_status, y=avg_price_per_room))+
  geom_point()+
  labs(title="Booking Status by Avergae Price per Room", x="Booking Staus", y="Avergae Price")+
  facet_wrap(~booking_status)
```

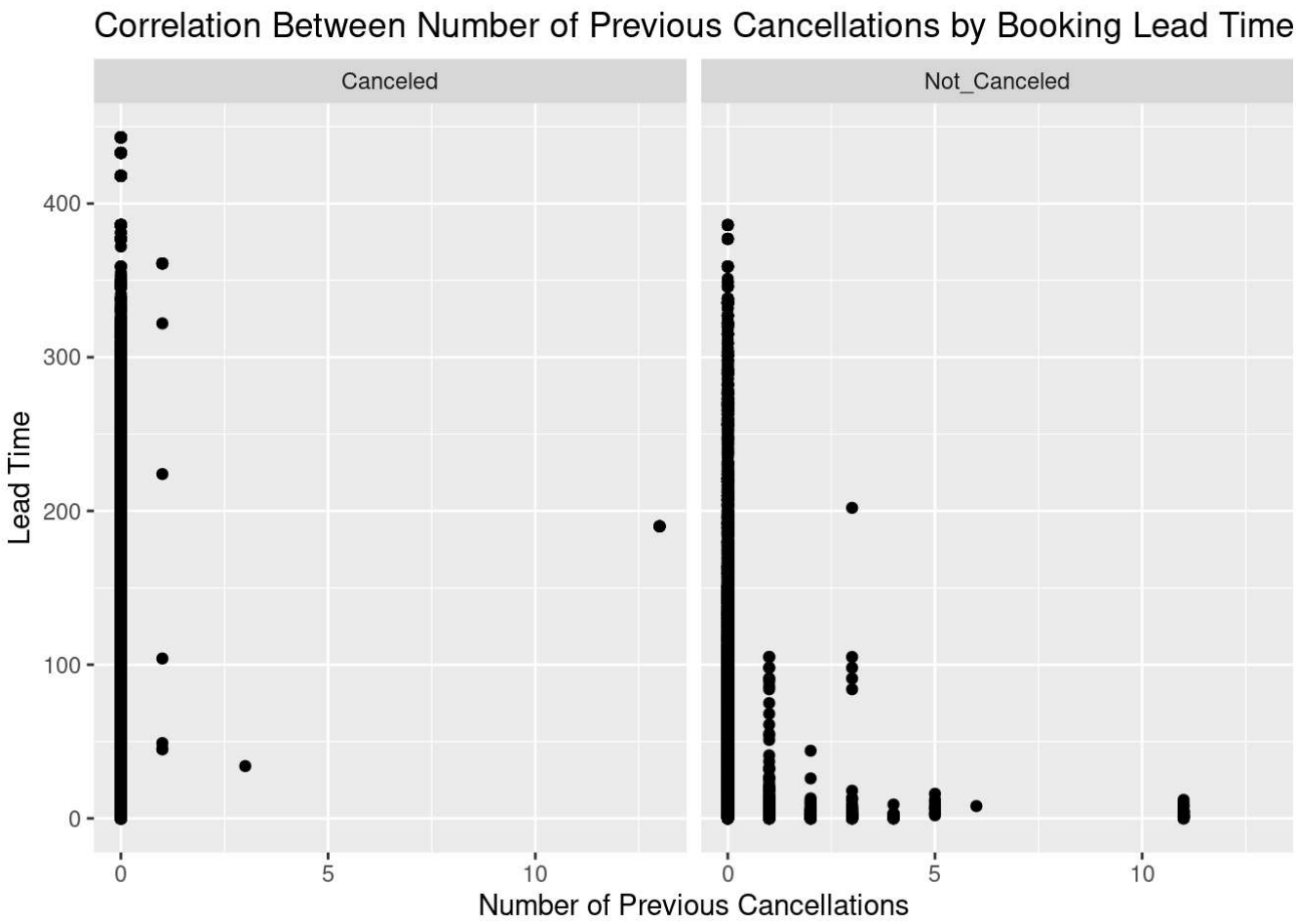


This graph above shows a even trend between the average price versus the booking being cancelled ot not except for an outlier showing once the price rises customers are more likely to cancel the booking. Let's see if one more factor can contribute to customers cancelling their booking.

```
ggplot(data, aes(x= no_of_children, fill=booking_status))+
  geom_bar()+
  labs(title= "Number of children by Booking Status", x="Number of Children", y="Frequency")+
  facet_wrap(~booking_status)
```




```
ggplot(data, aes(x =no_of_previous_cancellations, y = lead_time))+
    geom_point()+
    labs(title= "Correlation Between Number of Previous Cancellations by Booking Lead Time", x
="Number of Previous Cancellations", y="Lead Time")+
    facet_wrap(~booking_status)
```



After looking at various factors we can come to the conclusion that lead time and average price per room are the leading factors that contribute to a guest needing to cancel their stay. These tend to be typical reasons for cancellations but there is always other factors that contribute the the cancellation as well we can see that in the “correlation between number of previous cancellations plot that shows a good majority who have cancelled before are likely to do it again. Also with the plot showing where the booking was made that most online reservations are the ones to be cancelled more often then if it is a business trip or special occasion booking.