Hotel Booking Demand

Hatim Ali, Pavel Raschetnov

1/14/2021

Hotel booking demand dataset

This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details you can visit https://www.kaggle.com/jessemostipak/hotel-booking-demand.

```r
library(tidyverse)
```

Importing libraries

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

View Basic Attributes of Data

```r
hotel_bookings = read.csv("hotel_bookings.csv")
head(hotel_bookings)
```

1. View first 5 rows of data

```
##          hotel is_canceled lead_time arrival_date_year arrival_date_month
## 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
##   arrival_date_week_number arrival_date_day_of_month stays_in_weekend_nights
## 1                       27                         1                       0
## 2                       27                         1                       0
## 3                       27                         1                       0
## 4                       27                         1                       0
## 5                       27                         1                       0
## 6                       27                         1                       0
##   stays_in_week_nights adults children babies meal country market_segment
## 1                    0      2        0      0   BB     PRT         Direct
## 2                    0      2        0      0   BB     PRT         Direct
## 3                    1      1        0      0   BB     GBR         Direct
## 4                    1      1        0      0   BB     GBR      Corporate
## 5                    2      2        0      0   BB     GBR      Online TA
## 6                    2      2        0      0   BB     GBR      Online TA
##   distribution_channel is_repeated_guest previous_cancellations
## 1               Direct                 0                      0
## 2               Direct                 0                      0
## 3               Direct                 0                      0
## 4            Corporate                 0                      0
## 5                TA/TO                 0                      0
## 6                TA/TO                 0                      0
##   previous_bookings_not_canceled reserved_room_type assigned_room_type
## 1                             0                  C                  C
## 2                             0                  C                  C
## 3                             0                  A                  C
## 4                             0                  A                  A
## 5                             0                  A                  A
## 6                             0                  A                  A
##   booking_changes deposit_type agent company days_in_waiting_list customer_type
## 1               3   No Deposit  NULL    NULL                    0     Transient
## 2               4   No Deposit  NULL    NULL                    0     Transient
## 3               0   No Deposit  NULL    NULL                    0     Transient
## 4               0   No Deposit   304    NULL                    0     Transient
## 5               0   No Deposit   240    NULL                    0     Transient
## 6               0   No Deposit   240    NULL                    0     Transient
##   adr required_car_parking_spaces total_of_special_requests reservation_status
## 1   0                           0                         0          Check-Out
## 2   0                           0                         0          Check-Out
## 3  75                           0                         0          Check-Out
## 4  75                           0                         0          Check-Out
## 5  98                           0                         1          Check-Out
## 6  98                           0                         1          Check-Out
##   reservation_status_date
## 1              2015-07-01
## 2              2015-07-01
## 3              2015-07-02
## 4              2015-07-02
## 5              2015-07-03
## 6              2015-07-03
```

```
variables <- ncol(hotel_bookings)
rows <- nrow(hotel_bookings)
```

2. How many rows of data and how many variables? There are 32 variables with 119390 rows in this dataset. It looks like there are a lot of categorical variables in this dataset mixed with dates as well. An interesting metric they keep track of is number of special requests. Who knew hotels/resorts kept track of such things.

```
min_res_date <- min(hotel_bookings$reservation_status_date)
max_res_date <- max(hotel_bookings$reservation_status_date)
```

3. What is the data range for reservations? It appears that this data spans from 2014-10-17 to 2017-09-14.

```
glimpse(hotel_bookings)
```

4. Data type of each columns?

```
## Rows: 119,390
## Columns: 32
## $ hotel                          <chr> "Resort Hotel", "Resort Hotel", "Res...
## $ is_canceled                    <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, ...
## $ lead_time                      <int> 342, 737, 7, 13, 14, 14, 0, 9, 85, 7...
## $ arrival_date_year              <int> 2015, 2015, 2015, 2015, 2015, 2015, ...
## $ arrival_date_month             <chr> "July", "July", "July", "July", "Jul...
## $ arrival_date_week_number       <int> 27, 27, 27, 27, 27, 27, 27, 27, 27, ...
## $ arrival_date_day_of_month      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ stays_in_weekend_nights        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ stays_in_week_nights           <int> 0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, ...
## $ adults                         <int> 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, ...
## $ children                       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ babies                         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ meal                           <chr> "BB", "BB", "BB", "BB", "BB", "BB", ...
## $ country                        <chr> "PRT", "PRT", "GBR", "GBR", "GBR", "...
## $ market_segment                 <chr> "Direct", "Direct", "Direct", "Corpo...
## $ distribution_channel           <chr> "Direct", "Direct", "Direct", "Corpo...
## $ is_repeated_guest              <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ previous_cancellations         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ previous_bookings_not_canceled <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ reserved_room_type             <chr> "C", "C", "A", "A", "A", "A", "C", "...
## $ assigned_room_type             <chr> "C", "C", "C", "A", "A", "A", "C", "...
## $ booking_changes                <int> 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ deposit_type                   <chr> "No Deposit", "No Deposit", "No Depo...
## $ agent                          <chr> "NULL", "NULL", "NULL", "304", "240"...
## $ company                        <chr> "NULL", "NULL", "NULL", "NULL", "NUL...
## $ days_in_waiting_list           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ customer_type                  <chr> "Transient", "Transient", "Transient...
## $ adr                            <dbl> 0.00, 0.00, 75.00, 75.00, 98.00, 98....
## $ required_car_parking_spaces    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ total_of_special_requests      <int> 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, ...
## $ reservation_status             <chr> "Check-Out", "Check-Out", "Check-Out...
## $ reservation_status_date        <chr> "2015-07-01", "2015-07-01", "2015-07...
```

Data Wrangling

```
drop <- c("company","agent")
hotel_bookings = hotel_bookings[,!(names(hotel_bookings) %in% drop)]
```

Dropping the columns with missing values.

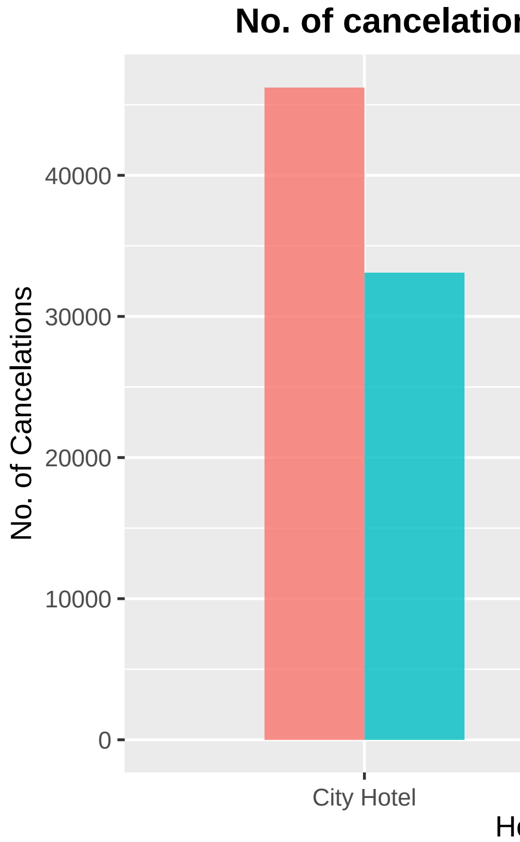
```
hotel_bookings$canceled <- hotel_bookings$is_canceled == 1
```

Adding canceled as a categorical variable

Data Visualizations (EDA):

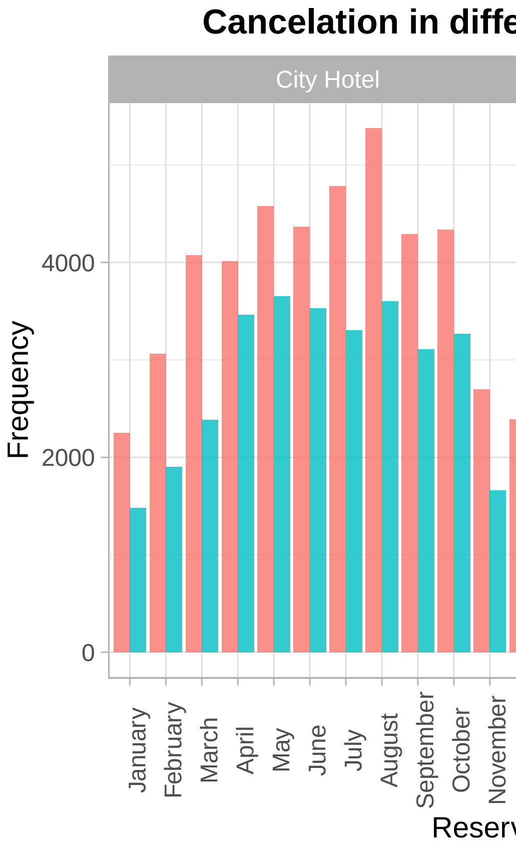
Lets explore the data with a couple of visualizations that will answer some interesting questions.

```
hotel_bookings %>%
  ggplot(aes(x= hotel, fill=canceled))+
  theme_set(theme_light()) +
  geom_bar(alpha=0.8, position = "dodge", width=0.5)+
  labs(title= "No. of cancelations with each hotel type", x= "Hotel Type", y="No. of Cancelations", fil
  theme(plot.title = element_text(face = "bold", hjust = 0.5), axis.text.x = element_text(vjust=.5))
```



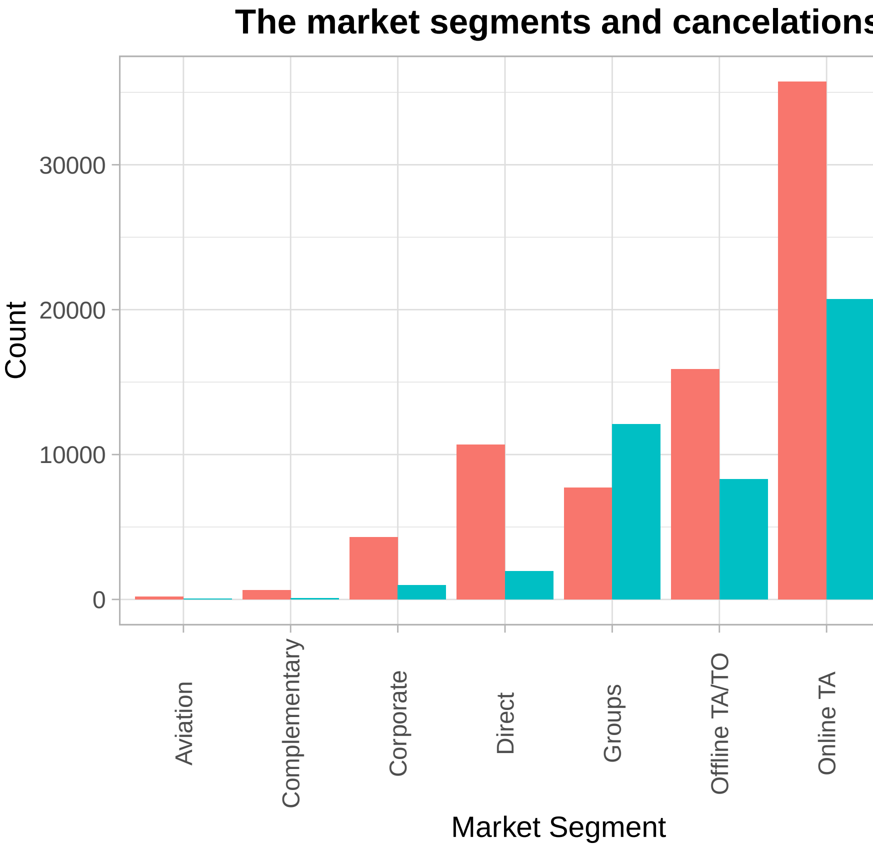
1. **What is the percentage of cancelled booking of each type of hotels?**

```
ggplot(hotel_bookings, aes(x = arrival_date_month, fill = canceled)) + geom_bar(position = "dodge", alpl
  scale_x_discrete(limits= month.name) +
  theme_set(theme_light()) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5), axis.text.x = element_text(angle=90, vju
  labs(title= "Cancelation in different months of the year", x='Reservation Month', y='Frequency') +
  facet_wrap(~hotel)
```



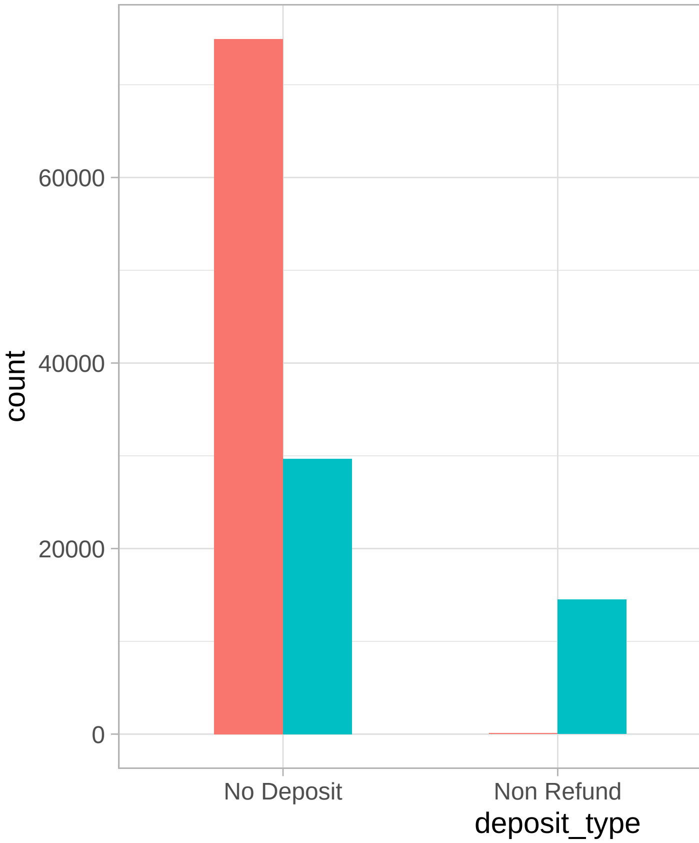
2. Distribution of the cancelation during different months of the year.

```
hotel_bookings$canceled <- hotel_bookings$is_canceled == 1
hotel_bookings %>%
  ggplot(aes(x = market_segment, fill = canceled)) +
  theme_set(theme_light()) +
  geom_bar(position = "dodge") +
  theme(plot.title = element_text(face = "bold", hjust = 0.5), axis.text.x = element_text(angle=90, vju
  labs(title= "The market segments and cancelations", x='Market Segment', y='Count')
```



3. The market segments and cancelations

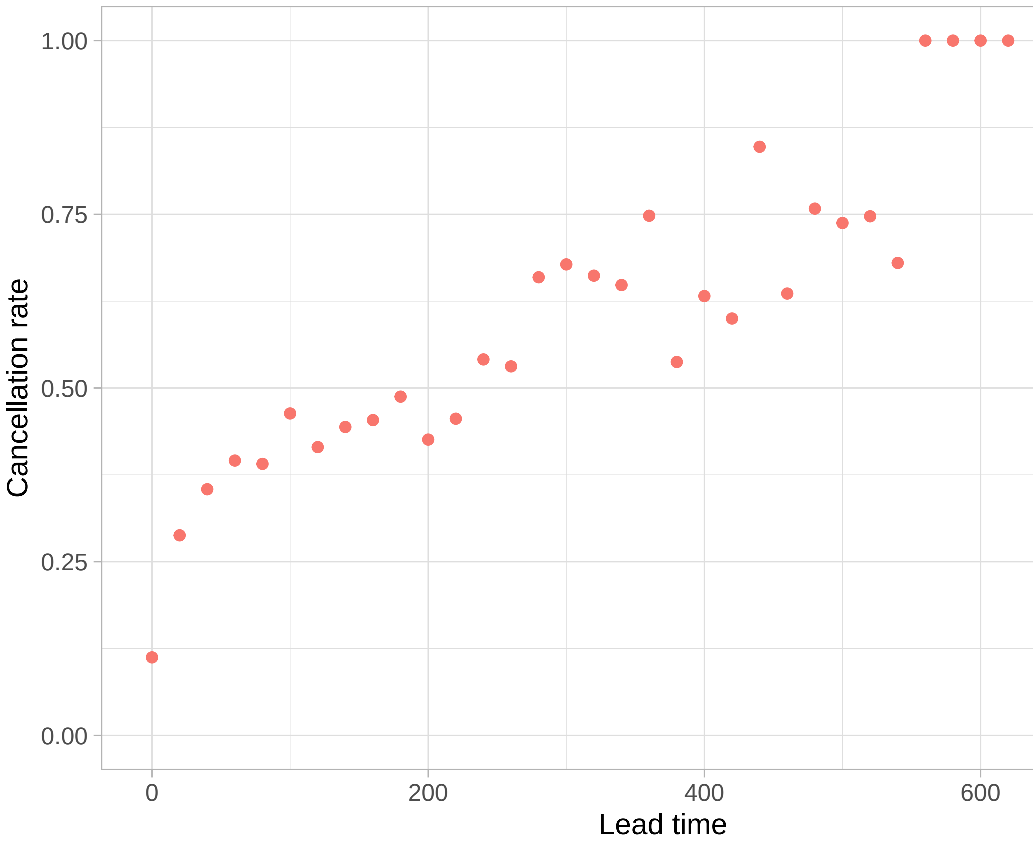
```
ggplot(hotel_bookings, aes(x = deposit_type, fill=canceled)) + geom_bar(position = "dodge", width=0.5) +
  theme_light() + scale_fill_discrete(name = "is_canceled", labels = c("confirmed", "canceled"))
```



4. Analyzing canceled booking based on deposit_type.

```
options(dplyr.summarise.inform = FALSE)
subset <- hotel_bookings  %>%
  mutate(lead_time_binned=round(lead_time / 20) * 20)  %>%
  group_by(lead_time_binned)  %>%
  summarise(cancellation_rate=mean(is_canceled)) %>%
  select(lead_time_binned, cancellation_rate)

ggplot(data= subset) +
  geom_point(aes(x=lead_time_binned, y=cancellation_rate, color='#eb5505'), show.legend = FALSE) +
  xlab('Lead time') + ylab('Cancellation rate')
```



5. Visualisation of lead_time.