Hotel Booking Cancellation

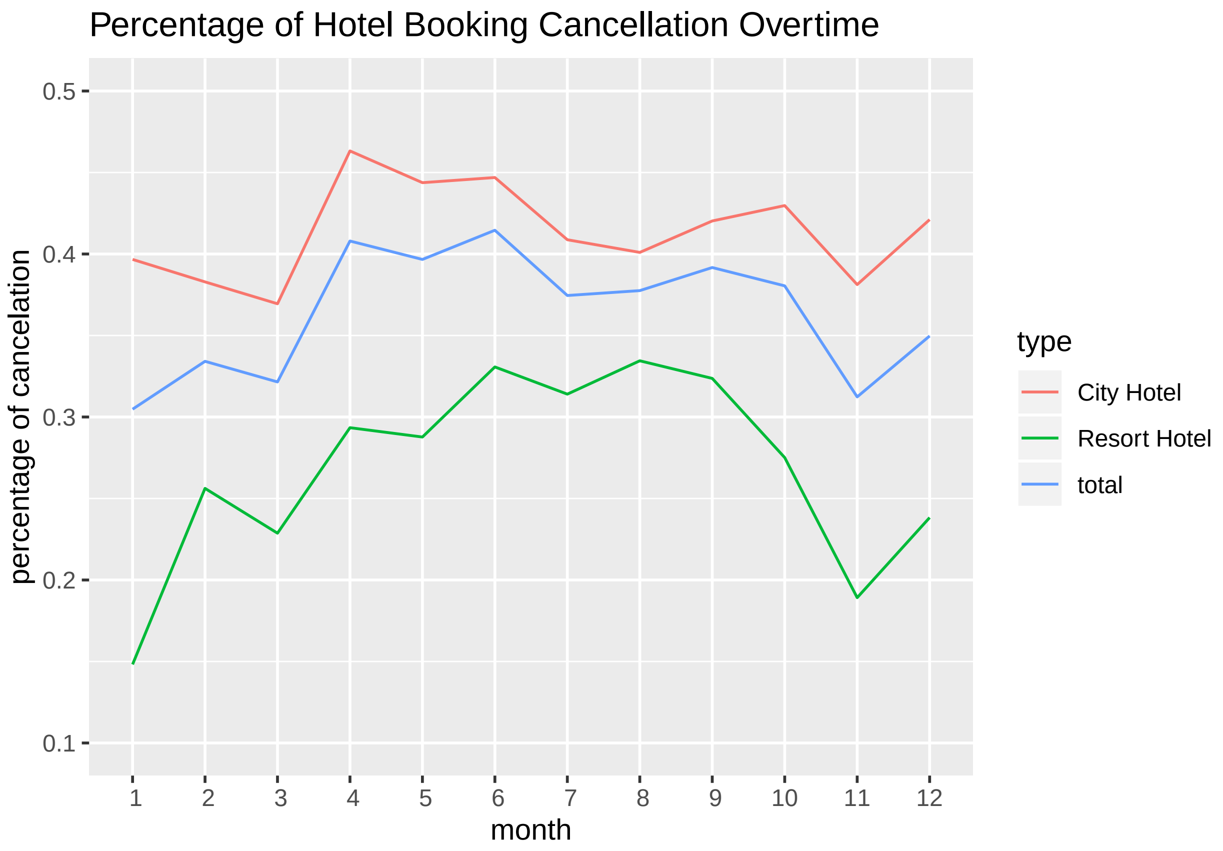
Shibi He

```
hotel = read.csv("hotel_bookings.csv", header = TRUE)
```

This project aims to identify important factors that influence the cancellation of hotel bookings and thereby provide recommendations for hotels to reduce cancellations and increase profits. The data I use is originally from the article “Hotel Booking Demand Datasets” written by Nuno Antonio, Ana de Almeida, and Luis Nunes (Data in Brief, Volume 22, February 2019). The dataset records booking information from a city hotel and a resort hotel between July 1, 2015 and Augst 31, 2017 and contains information such as whether the booking was canceled, when the booking was made, the type of deposit, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. Before developing the cancellation predication model, I did some exploratory data analysis. In the first plot I show the percentage of cancellation by month for the city hotel and the resort hotel separately. I found (i) the percentage of cancellation of the city hotel is much higher than the resort hotel, and (ii) there is a sharp decline during the holiday season (November to January). In the second plot I show the percentage of cancellation for different types of deposit. Surprisingly, I found bookings with non-refundable bookings have extremely high chance of cancellation while bookigs without deposit have the lowest chance of cancellation. I believe this has something to do with the hotel’s daily rate and need further exploration.

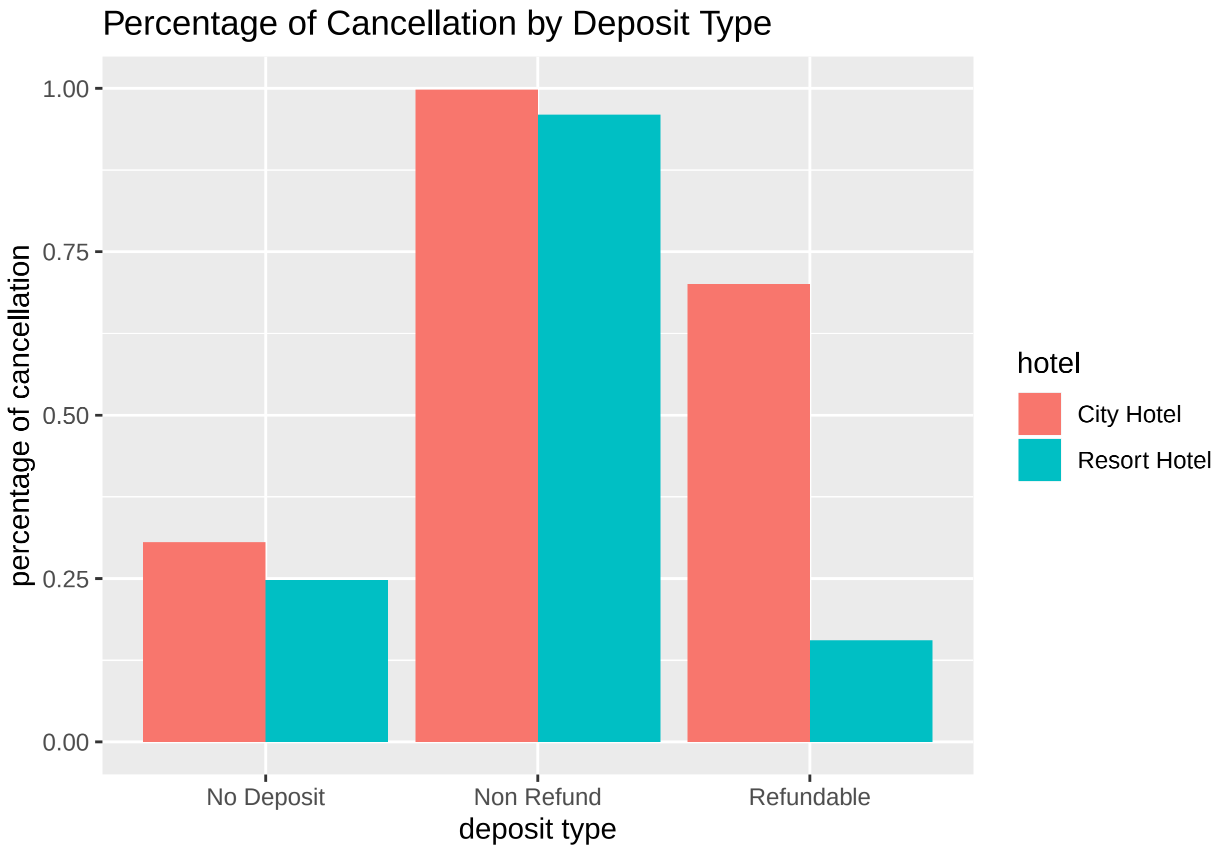
```
# percentage of canceled vs month
summary = hotel %>%
  group_by(arrival_date_month) %>%
  summarize(Ncanceled = sum(is_canceled), count = n(), p_canceled = Ncanceled/count)

summary2 = hotel %>%
    group_by(arrival_date_month, hotel) %>%
   summarize(Ncanceled = sum(is_canceled), count = n(), p_canceled = Ncanceled/count)

df = subset(summary2, select = -c(Ncanceled,count) )
summary2_wide = spread(df, hotel, p_canceled)
summary2_wide$total = summary$p_canceled
summary2_wide$month = 4
summary2_wide$month[2] = 8
summary2_wide$month[3] = 12
summary2_wide$month[4] = 2
summary2_wide$month[5] = 1
summary2_wide$month[6] = 7
summary2_wide$month[7] = 6
summary2_wide$month[8] = 3
summary2_wide$month[9] = 5
summary2_wide$month[10] = 11
summary2_wide$month[11] = 10
summary2_wide$month[12] = 9

summary2_long = gather(summary2_wide, type, p_cancel, "City Hotel":total, factor_key=TRUE)
```

```
# plot
ggplot(summary2_long, aes(x=month, y=p_cancel, color=type))  + geom_line() + scale_x_discrete(limits=c(
```



```
data2 = hotel %>%
    group_by(deposit_type, hotel) %>%
   summarize(Ncanceled = sum(is_canceled), count = n(), p_canceled = Ncanceled/count)
```

```
ggplot(data2, aes(fill=hotel, y=p_canceled, x=deposit_type)) +
    geom_bar(position="dodge", stat="identity")+ ggtitle("Percentage of Cancellation by Deposit Type") +
```


Percentage of Cancellation by Deposit Type
percentage of cancellation
1.00
0.75
0.50
0.25
0.00
No Deposit
Non Refund
Refundable
deposit type
hotel
City Hotel
Resort Hotel