36-402 Homework 10

James “Morgan” Hawkins

Friday April 21, 2023

Problem 1

```
hotels = read.csv("/Users/morganhawkins/Downloads/hotels.csv")
colnames(hotels)
```

```
## [1] "no_of_adults"         "no_of_children"       "no_of_weekend_nights"
## [4] "no_of_week_nights"    "lead_time"            "avg_price_per_room"
## [7] "canceled"
```

```
cat(dim(hotels))
```

```
## 1231 7
```

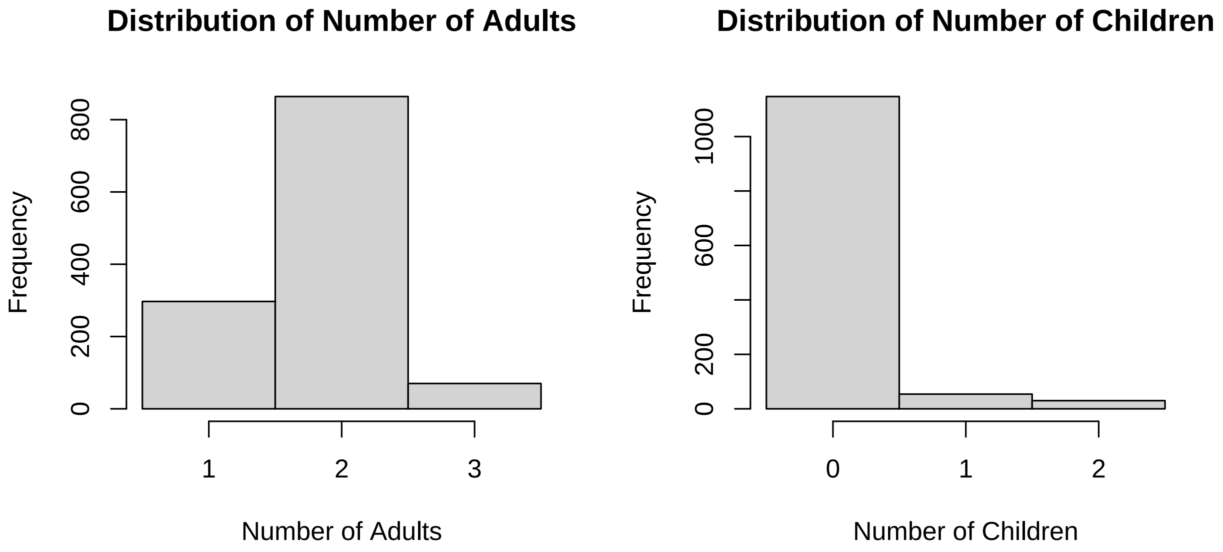
Problem 1 (a)

```
table(hotels$canceled)
```

```
##
##   0   1
## 822 409
```

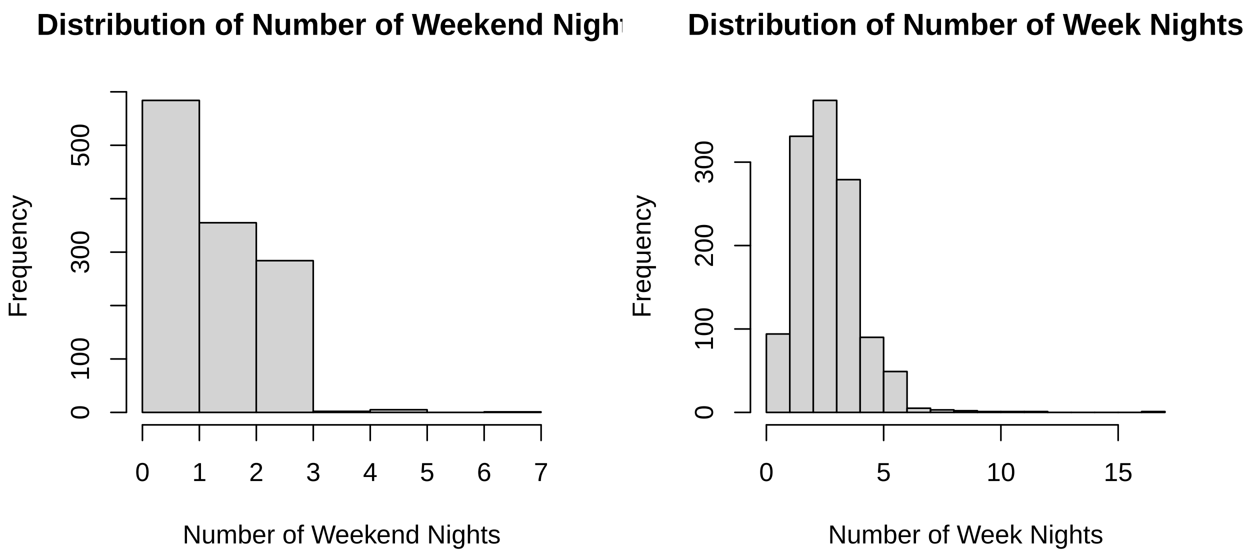
In our data set we see that our response variable shows 409 canceled reservations (33.2%) and 822 not canceled reservations (66.8%). So, cancellations are quite common at the hotel where this data was collected.

```
#colnames(hotels)
#table(hotels$no_of_adults)/length(hotels$no_of_adults)
#table(hotels$no_of_children)/length(hotels$no_of_adults)

par(mfrow = c(1,2))
hist(hotels$no_of_adults,
     main = "Distribution of Number of Adults",
     xlab = "Number of Adults",
     breaks = 4, xaxt = 'n')
axis(1, at = c(.5,1.5,2.5), labels = c(1,2,3))

hist(hotels$no_of_children,
     main = "Distribution of Number of Children",
     xlab = "Number of Children",
     breaks = (seq(0,3,1) - .001),
     xaxt = 'n')
axis(1, at = c(.5,1.5,2.5), labels = c(0,1,2))
```



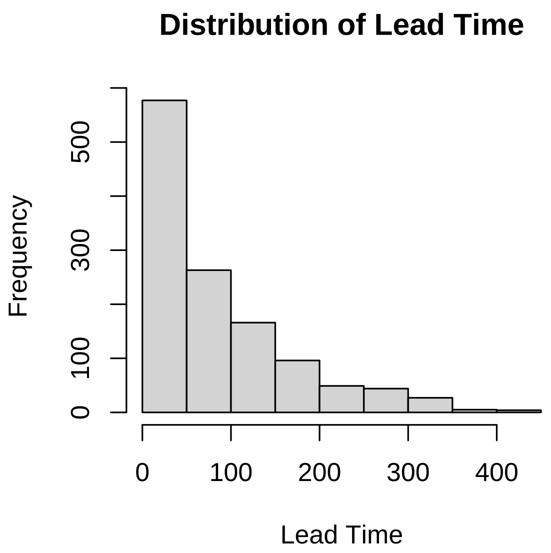
Above we see that most reservations have 2 adults (70.2%), 1 adult is the second most common (23.8%), and 3 adults is the least common (5.7%). There are 4 reservations with 0 adults (.3%) on the reservation, but we questions whether this is an error in data reporting or if the hotel allows minors to make hotel reservations. We also see that the vast majority of reservations have 0 children on them (93.2%). The second most common is 1 child (4.4%) and the least common is 2 children (2.4%).

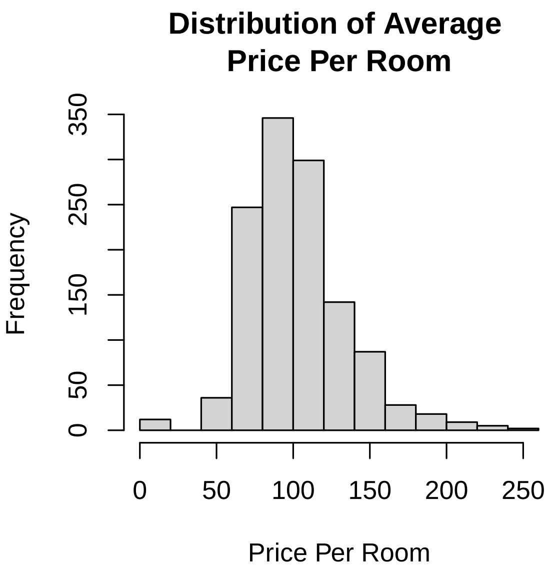
```
#table(hotels$no_of_weekend_nights)/length(hotels$no_of_weekend_nights)
#table(hotels$no_of_week_nights)/length(hotels$no_of_week_nights)


par(mfrow = c(1,2))
hist(hotels$no_of_weekend_nights,
     main = "Distribution of Number of Weekend Nights",
     xlab = "Number of Weekend Nights",
     breaks = (seq(0,7,1) - .001))
hist(hotels$no_of_week_nights,
     main = "Distribution of Number of Week Nights",
     xlab = "Number of Week Nights",
     breaks = (seq(0,17,1) - .001))
```



Above we see that the most common number of weekend nights for a reservation to have is 0 nights (47.4%). The second most common number of weekend nights is 1 night (28.8%) and the third most common is 2 nights (23.1%). We notice that this sums to almost 100% and this is not surprising because stays must be at least 8 days to have more than 2 weekend day stays. Next we see that the most common number of week days on a reservation is 2 (30.4%). Other common numbers are 1 day (26.9%) and 3 days (22.7%). Reservations with 1,2, or 3 week days ont he reservation represent more than half of reservations made.

```r
#mean(hotels$lead_time)
#sd(hotels$lead_time)

#mean(hotels$avg_price_per_room)
#sd(hotels$avg_price_per_room)

#filter(hotels, avg_price_per_room< 37)


par(mfrow = c(1,2))
hist(hotels$lead_time,
     main = "Distribution of Lead Time",
     xlab = "Lead Time")
hist(hotels$avg_price_per_room,
     main = "Distribution of Average \nPrice Per Room",
     xlab = "Price Per Room")
```





Above we see that the distribution of lead time seems to have monotonically decreasing density. The distribution is unimodal and heavily right skewed. The mean lead time for reservations is 82.4 days and the standard deviation is 84.2. The distribution of price per night appears somewhat normal although right skewed. The average price per night is $103.65 and the standard deviation is 32.89. There are 12 outliers in price per night. 10 of these reservations have price per night values of $0 and the other two are $1 and $12. This may be an error in data reporting, but we cannot be certain.

```
filter(hotels, no_of_adults>0)[,c("no_of_adults","canceled")] %>%
  table()/cbind(unname(table(hotels$no_of_adults)[-1]), unname(table(hotels$no_of_adults)[-1]))
```

```
##             canceled
## no_of_adults         0         1
##            1 0.7372014 0.2627986
##            2 0.6481481 0.3518519
##            3 0.6000000 0.4000000
```

We would now like to explore the relationships between our explanatory variables and our response. Beginning with number of adults on the reservation we see that compared to the overall cancellation frequency of 33.2%, reservations with 1 adult are less likely to cancel while reservations with 2 or 3 adults are more likely to cancel. This is shown numerically in the table above.

```
#lm(canceled ~ factor(no_of_children), data = hotels) %>% summary

filter(hotels)[,c("no_of_children","canceled")] %>%
  table()/cbind(unname(table(hotels$no_of_children)),
                unname(table(hotels$no_of_children)))
```

```
##               canceled
## no_of_children         0         1
##              0 0.6678291 0.3321709
##              1 0.6851852 0.3148148
##              2 0.6333333 0.3666667
```

Next, we see that number of children doesn't appear to have a large effect on cancellation frequency. This is reinforced by an f test conducted through a linear model predicting canceled from number of children ($F_{2,1228}$ = .1168, p = .8897).

```
#lm(canceled ~ factor(no_of_weekend_nights), data = hotels) %>% summary

filter(hotels)[,c("no_of_weekend_nights","canceled")] %>%
  table()/cbind(unname(table(hotels$no_of_weekend_nights)),
                unname(table(hotels$no_of_weekend_nights)))
```

```
##                       canceled
## no_of_weekend_nights         0         1
##                    0 0.7020548 0.2979452
##                    1 0.6450704 0.3549296
##                    2 0.6338028 0.3661972
##                    3 0.5000000 0.5000000
##                    4 0.4000000 0.6000000
##                    6 0.0000000 1.0000000
```

In the table above we notice that an increase in the number of weekend nights stayed a the hotel appears to be associated with an increase in cancellation frequency. However, an F test conducted through a linear model predicting canceled from number of weekend nights does not find this relationship significant ($F_{5,1225}$ = 1.86, p = .0985).

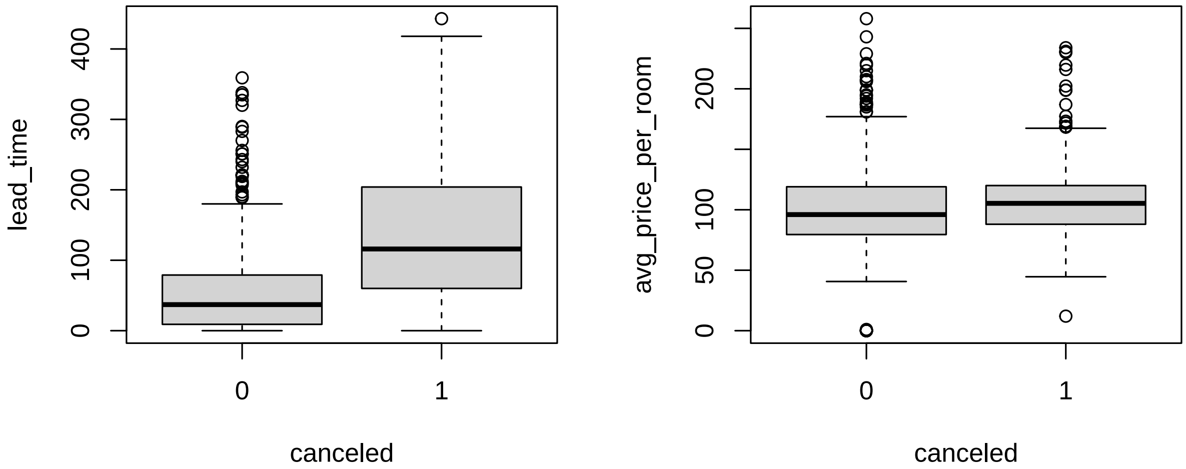
```
#lm(canceled ~ no_of_week_nights, data = hotels) %>% summary

filter(hotels)[,c("no_of_week_nights","canceled")] %>%
  table()/cbind(unname(table(hotels$no_of_week_nights)), unname(table(hotels$no_of_week_nights)))
```

```
##                    canceled
## no_of_week_nights         0         1
##                 0  0.6808511 0.3191489
##                 1  0.7371601 0.2628399
##                 2  0.6604278 0.3395722
##                 3  0.6666667 0.3333333
##                 4  0.5777778 0.4222222
##                 5  0.4897959 0.5102041
##                 6  0.6000000 0.4000000
##                 7  0.3333333 0.6666667
##                 8  0.5000000 0.5000000
##                 9  0.0000000 1.0000000
##                 10 0.0000000 1.0000000
##                 11 0.0000000 1.0000000
##                 16 0.0000000 1.0000000
```

In the table and plot above we notice that an increase in the number of week nights stayed a the hotel appears to be associated with an increase in cancellation frequency. An F test conducted through a linear model predicting canceled from number of weekend nights did find this relationship significant ($F_{1,1229}$ = 19.93, p = 8.756e-6).

```
#lm(avg_price_per_room ~ canceled, data = hotels) %>% summary
```

```
par(mfrow = c(1,2))

boxplot(lead_time ~ canceled, data = hotels)
boxplot(avg_price_per_room ~ canceled, data = hotels)
```



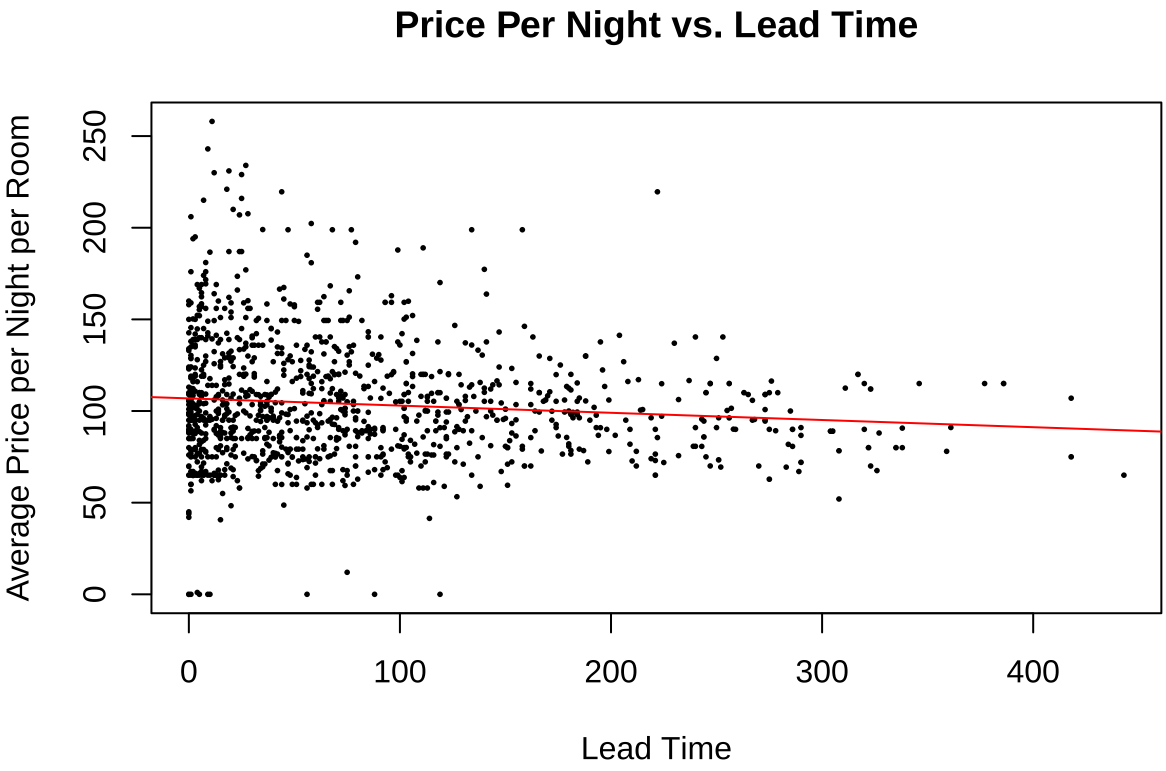
In the first box plot above (lead time vs canceled), we notice that canceled reservations have much higher median lead times. We also notice that reservations that weren't canceled have far more outliers in lead time. In the second box plot we see that canceled and not canceled reservations have similar distributions of average price per night. However, canceled reservations have a slightly higher average price per night. This observed relationship was found to be significant through a F test on a linear model predicting average room price per night from canceled ($F_{1,1229}$ = 8.319, p = 3.991e-3).

```
#lm(no_of_children ~ no_of_adults == 2, data = filter(hotels, no_of_adults > 0)) %>% summary

cor(hotels[,-(5:7)])%>% round(digits = 3) %>%
  kable() %>%
  kable_styling(position = "left",html_font = "helvetica", font_size = 10,
              full_width = F) %>%
  row_spec(0, angle = 45, font_size = 7)
```

	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights
no_of_adults	1.000	-0.007	0.112	0.098
no_of_children	-0.007	1.000	-0.001	0.032
no_of_weekend_nights	0.112	-0.001	1.000	0.151
no_of_week_nights	0.098	0.032	0.151	1.000

Now we would like to explore the correlation between our discrete predictor variables. We notice correlation between the number of weekend nights and the number of week nights as well as correlation between the number of adults and number of weekend nights. The number of adults and children on the reservation does not appear to be linear correlated when we consider number of adults as a continuous variable. However, when we predict number of children from (number of adults $= 2$) we see that there is a significant relationship as shown by a t test ($T_{1225} = 4.039, p = 0.0000571$).

```
plot(avg_price_per_room ~ lead_time, data = hotels,
     pch = 19, cex = .25,
     xlab = "Lead Time",
     ylab = "Average Price per Night per Room",
     main = "Price Per Night vs. Lead Time")
lm(avg_price_per_room ~ lead_time, data = hotels) %>% abline(col = 'red')
```



Lastly, we would like to explore the relationship between our continuous predictor variables. We that there is a negative approximately linear relationship between lead time and price per night. We also notice that there is a lot more variance in price per night for smaller values of lead time.

From this exploratory data analysis we suspect that the variables `lead_time` and `no_of_weeknights` with be the strongest predictor of our response variable `canceled`