Predicting the Cost of Hotel Booking in the United States

LAM-duh: Xuliang Deng, Leah Okamura, Megan Richards

Nov 15th, 2021

Introduction

With the lifting of travel restrictions into the U.S. (https://www.nytimes.com/2021/09/22/travel/us-international-travel-vaccine.html) through the implementation of new travel guidelines, we believe that travel to the U.S. start to increase and look like as it did before the pandemic. Therefore, with travel and tourism in the U.S. increasing, we believe that the booking of hotels may start to increase as well. With this assumption, with the slower return to travel and society pre covid, we are interested in studying the characteristics of hotel room reservations in the United States. Specifically, we are interested in what relationship these characteristics have with the cost of a hotel. Our general research question is: How do the characteristics of a hotel booking affect the daily cost of a hotel stay in the United States? We believe there will be several significant points of relevance for understanding these relationships: understanding predictors of room cost could be used to help identify where new hotels could be successfully created, allow travelers to plan financially for future travel, allow hotels to predict future profit, etc. In this report, we are looking to use a variety of chosen models to understand the contributing factors to hotel room price, as well as identify the strongest predictors in the determining the avreage daily rate.

Data

The source of the dataset is Tiny Tuesday, https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-02-11/readme.md. This data set was originally collected from an open hotel booking demand dataset from Antonio, Almeida and Nunes, 2019. The data collected from hotels all around the world ranges from bookings in 2015 to 2017. It is sourced from this study https://www.sciencedirect.com/science/article/pii/S2352340918315191#f0010. Due to the dataset being over 100,000 observations, we have limited the observations to be only hotels from the US. The general characteristics being measured in the data are the different aspects of booking and staying at a hotel. For example, out of the 32 variables, some of the ones we find great interest in are hotel type, reserved room type, assigned room type, company, meal, number of adults/children/babies, the average daily rate or daily cost, and the reservation status. Therefore, each observation is one booking/stay at a specific hotel in America and all of its characteristics. Therefore, there can be multiple observations from the same hotel and even on the same time range. The full dataset dictionary can be found in the repository ReadMe.

Resesarch Question

To better understand the contributing factors to the daily rate of hotel rooms in the U.S., we will build models to:

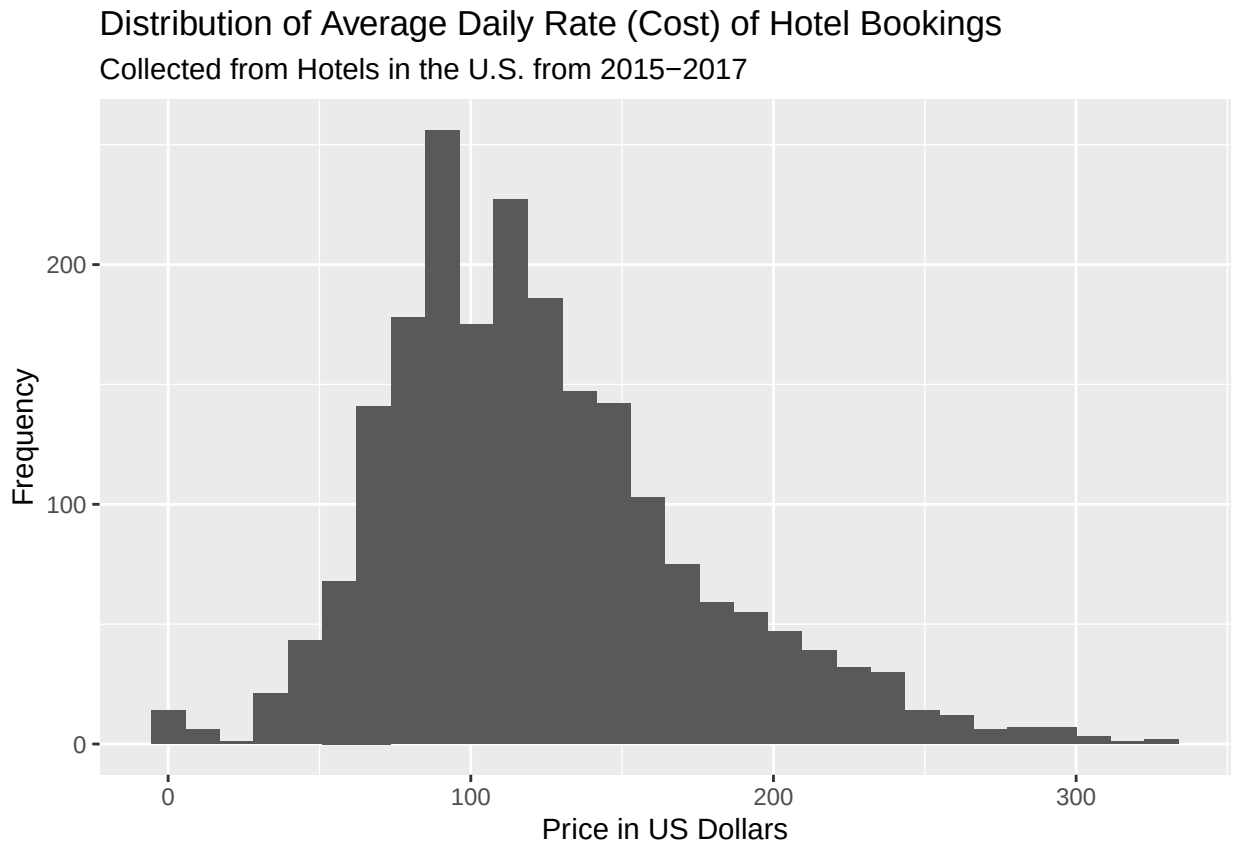
1) Evaluate the efficacy of this model to predict daily hotel room rates
2) Evaluate the statistical significance of each predictor, identifying the strongest contributing factors. We include the following variables in our analysis: hotel type, company, meal, number of adults/children/babies, the average daily rate or daily cost, and the reservation status.

Exploratory data analysis

```
hotels$total_stay_length <- hotels$stays_in_week_nights + hotels$stays_in_weekend_nights
```

```
ggplot(data = hotels, aes(x = adr)) +
  geom_histogram() +
  labs(x = "Price in US Dollars",
       y = "Frequency",
       title = "Distribution of Average Daily Rate (Cost) of Hotel Bookings",
       subtitle = "Collected from Hotels in the U.S. from 2015-2017")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
hotels %>%
    summarise(mean = mean(adr),
              median = median(adr),
              sd = sd(adr),
              min = min(adr),
              max = max(adr),
              iqr = IQR(adr))%>%
kable(digits = 3)
```

mean	median	sd	min	max	iqr
122.992	115	51.617	0	328.33	61.99

The response variable, adr, has a somewhat skewed right, bimodal distribution. The average or mean average daily rate is $122.992 and the median is $115. Because the distribution is skewed, the median is most likely the best indicator for the center. The standard deviation is $51.617 and the data ranges from $0 to $328.33 with an interquartile range of $61.99.

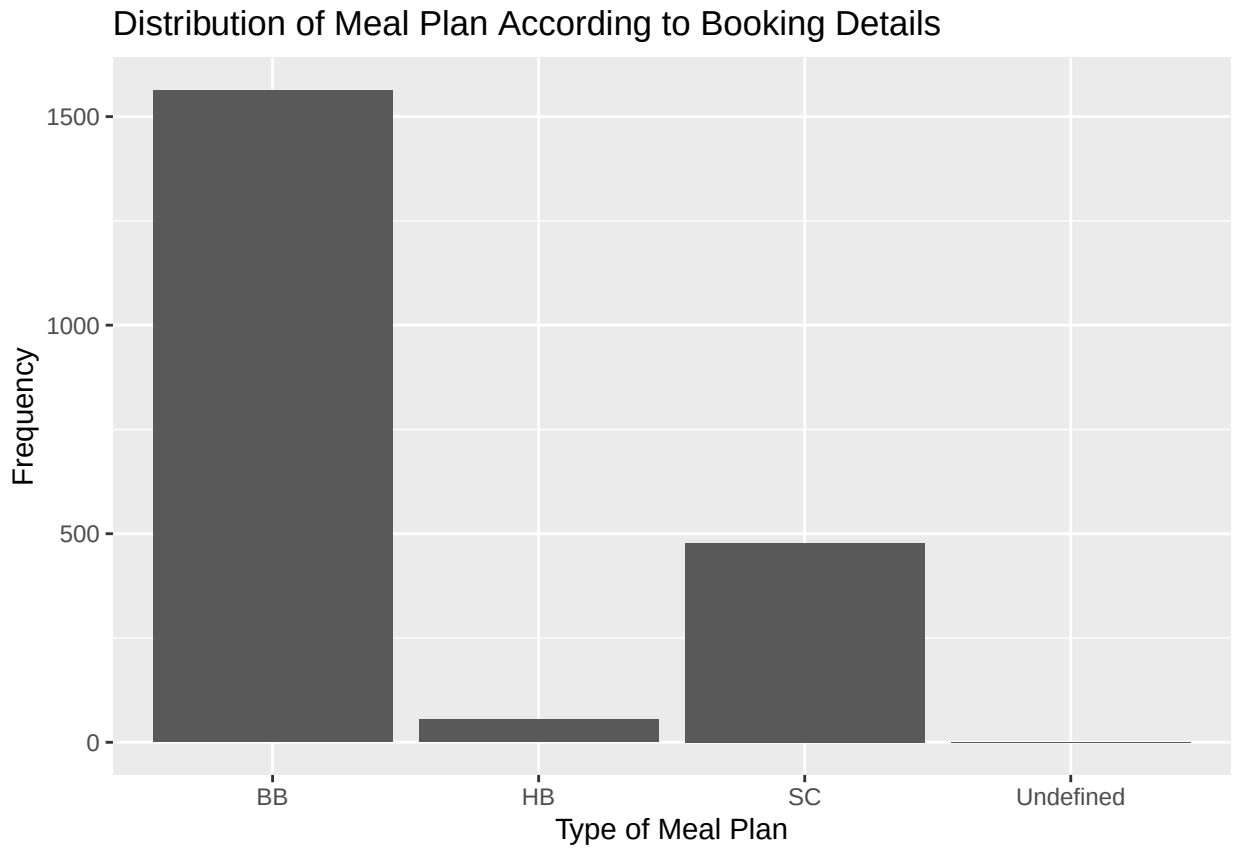
```
adults <- sum(hotels$adults)
children <- sum(hotels$children)
babies <- sum(hotels$babies)

tb <- table(adults, children, babies)
kable(tb)
```

adults	children	babies	Freq
3950	362	6	1

This table provides insight on the type of people who were booking hotels in the U.S. during this time period. A large majority of customers were only adults with a total of 3,950 hotel bookings. The next largest customer group were people or adults with their children at a total of 362 bookings. Lastly, those who booked a stay with a baby was considerably less, with only 6 bookings.

```
ggplot(data = hotels, aes(x = meal)) +
geom_bar() +
labs(title = "Distribution of Meal Plan According to Booking Details",
      y = "Frequency",
      x = "Type of Meal Plan")
```

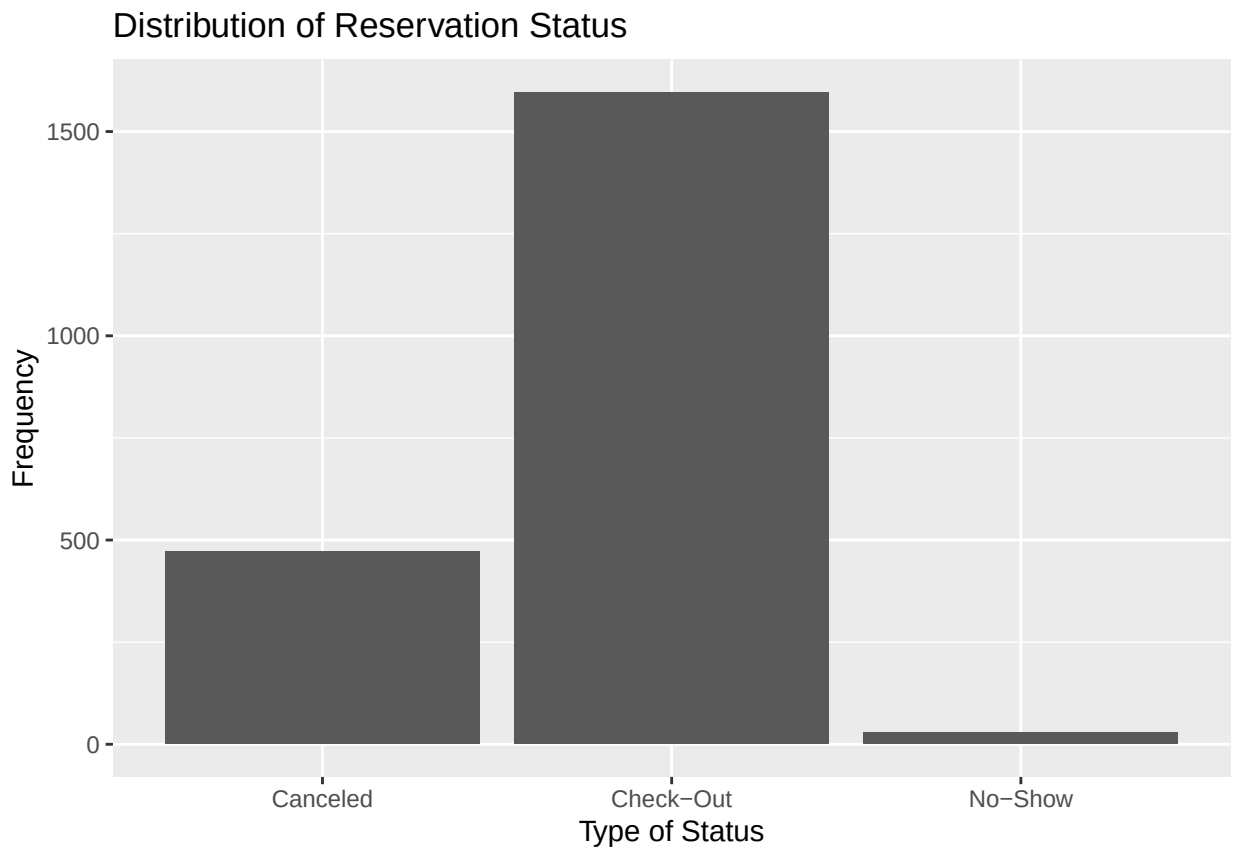



```
hotels %>%
  group_by(meal) %>%
  summarise(n = n(), mean = mean(adr), sd = sd(adr)) %>%
  kable(digits = 3)
```

meal	n	mean	sd
BB	1563	127.834	54.168
HB	55	167.216	60.824
SC	478	101.678	29.128
Undefined	1	311.000	NA

We can see that majority of customer bookings chose the Bed and Breakfast plan (BB), rather than Half Board plan or Full Board (HB and SC). While this does reflect the choice of those booking the hotel, this visualization and dataset can also be influenced by what type of meal plan each hotel is offering.

```
ggplot(data = hotels, aes(x = reservation_status)) +
geom_bar() +
labs(title = "Distribution of Reservation Status",
  x = "Type of Status",
  y = "Frequency")
```



A vast majority of the reservation statuses in this dataset resulted in a check-out, meaning that the customer arrived and stayed for the duration of the stay. It also important to note that a substantial proportion of hotel bookings were canceled, and if possible we would like to investigate what relationship might exist here.

```
hotels %>%
  group_by(hotel) %>%
  summarise(n = n(), mean = mean(adr), sd = sd(adr)) %>%
  kable(digits = 3)
```

hotel	n	mean	sd
City Hotel	1618	119.773	45.189
Resort Hotel	479	133.865	67.982

We can see that resort hotel is on average more expensive, and that most customers booked at city hotels. We are interested in whether the classification of City Hotel and Resort Hotel will proportionately have similar relationships with the other variables.

```
ggplot(data = hotels, aes(x = lead_time, y = adr, color = hotel)) +
geom_point() +
labs(title = "Relationship between how early people book hotel and price in the US dollars",
  x = "Subtraction of entering date from arrival date",
  y = "Price in the US Dollars")
```

We can see that the more last minute visitors book a hotel, the more likely it is that the price in the US dollars varies. Besides, we can also see that the highest price in a hotel decreases as the difference between entering data and arrival date increases. This could suggest that the earlier you book a hotel, the price is more likely to be cheap.

```
ggplot(data = hotels, aes(x = adr, y = arrival_date_day_of_month, color=hotel)) +
geom_point() +
labs(title = "Relationship between day in the month when hotels are booked and price in the US dollars"
  x = "Day in the month",
  y = "Price in US Dollars")
```

We cannot see any significant relationship between day in the month and the hotel price from this figure. However, we can see that summer is the busiest season for hotels and the price varies evenly. At the beginning of the year and the end of the year, hotels are not really booked as often.

```
ggplot(data = hotels, aes(x = arrival_date_week_number, y = adr, color=hotel)) +
geom_point() +
labs(title = "Relationship between week in the year when hotels are booked and price in the US dollars"
  x = "Week in the year",
  y = "Price in US Dollars")
```

From this figure, we can also see that at the beginning and end of the year hotels are not as busy as during the summer. Besides, we can also observe that the price of hotel tends to become higher during summer time compared to the rest of the year.

We are particularly interested in using the hotel type (Resort versus City Hotel) as a predictor of average daily rate. As an initial analysis, we perform a t-test on the the average daily rates of bookings for resort versus city hotels. This t-test assesses if there is a significant difference between the two distributions, which would indicate that hotel type is likely to be a useful predictor of average daily rate. In this analysis, we find that our t-test produces a p-value of 0.002305%. Because this p-value is very low, we conclude that there is a statistically significant relationship between the average daily hotel rates of resort hotels and city hotels in our dataset, which provides evidence that hotel type is a useful predictor of average daily hotel rates.

```
resort_adr <- hotels[hotels$hotel == "Resort Hotel",]$adr
city_adr <- hotels[hotels$hotel == "City Hotel",]$adr

hotel_type_p_value <- t.test(resort_adr, city_adr)$p.value
print(hotel_type_p_value)
```

```
## [1] 2.304792e-05
```

We are also particularly interested in using the presence of kids (children or babies) as a predictor of average daily rate. As an initial analysis, we perform a t-test on the the average daily rates of bookings for groups with only adults, versus groups with children or babies. This t-test assesses if there is a significant difference between the two distributions, which would indicate that the presence of kids is likely to be a useful predictor of average daily rate. In this analysis, we find that our t-test produces a p-value of 1.13e-33%. Because this p-value is very low, we conclude that there is a statistically significant relationship between the average daily

hotel rates of groups with children or babies and groups with only adults in our dataset, which provides evidence that the presence of kids (children or babies) is a useful predictor of average daily hotel rates.

```
hotels$only_adults_present <- hotels$adults > 0 & hotels$children == 0 & hotels$babies == 0
hotels$kids_present <- hotels$children > 0 | hotels$babies > 0

adult_adr <- hotels[hotels$only_adults_present == TRUE,]$adr
kids_adr <- hotels[hotels$kids_present == TRUE,]$adr

resident_type_p_value <- t.test(adult_adr, kids_adr)$p.value
print(resident_type_p_value)
```

```
## [1] 1.134618e-35
```

Methodology

Modeling Choices (Description)

- Data transformations (if applied)
- Model Type Justification
- Model selection criteria
- Handling Interaction Terms
- Any other technical choices

It is important to note that there are a few variables in the dataset that do not provide insight to our research question and as a result we will remove from the dataset. First is the category `hotels$country`, due to the fact that every hotel in this subset of the data is located in the U.S., this is a redundant variable. Next, the two variables `hotels$reserved_room_type` and `hotels$assigned_room_type` are both categorical variables that have letters assigned to each room type. However, due to the confidentiality of the customers, the classification of these letters have not been identified by those who collected the data. Therefore, we will be removing these variables from the dataset.

```
hotels$country <- NULL
hotels$reserved_room_type <- NULL
hotels$assigned_room_type <- NULL
```

As of now, we know we are going to use hotel (Resort Hotel or City Hotel), company, customer_type, stays_in_weekend_nights, stays_in_week_nights, and meal(type of meal) as predictor variables. We are interested in how the food is served at a hotel and what type of hotel can indicate how much a night in a hotel could cost. As we explore more of the relationships in our dataset, we may add other possible predictor variables as we see fit.

We plan to use multiple linear regression with a variety of combinations of interested predictor variables and their interaction. Based on whether conditions or fit or not or based on the visualization, we believe we may be utilizing logarithmic regression as well. We would also be interested in seeing how stays in the weekend or the weekday may affect the average daily rate for a hotel, and if they differ between the two hotel types, City and Resort hotels.

```
full_model <- lm(adr ~ adults + children + babies + hotel + meal + reservation_status, data = hotels)
tidy(full_model) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	79.723	4.256	18.734	0.000
adults	26.277	1.825	14.398	0.000
children	35.169	1.825	19.274	0.000
babies	-22.637	17.815	-1.271	0.204
hotelResort Hotel	2.353	2.391	0.984	0.325
mealHB	33.779	5.982	5.647	0.000
mealSC	-17.205	2.409	-7.141	0.000
mealUndefined	106.033	43.632	2.430	0.015
reservation_statusCheck-Out	-12.809	2.323	-5.515	0.000
reservation_statusNo-Show	-2.662	8.327	-0.320	0.749

```
selected_model <- step(full_model, scope=formula(full_model), direction="backward", k = log(nrow(hotels)
```

```
## Start:  AIC=15883.75
## adr ~ adults + children + babies + hotel + meal + reservation_status
##
##                      Df Sum of Sq     RSS   AIC
## - hotel               1      1828 3940298 15877
## - babies              1      3047 3941517 15878
## <none>                             3938470 15884
## - reservation_status  2     58774 3997244 15900
## - meal                3    176742 4115212 15953
## - adults              1    391223 4329693 16075
## - children            1    701016 4639486 16220
##
## Step:  AIC=15877.08
## adr ~ adults + children + babies + meal + reservation_status
##
##                      Df Sum of Sq     RSS   AIC
## - babies              1      2815 3943113 15871
## <none>                             3940298 15877
## - reservation_status  2     57316 3997614 15892
## - meal                3    197713 4138011 15957
## - adults              1    396718 4337016 16071
## - children            1    707852 4648149 16216
##
## Step:  AIC=15870.93
## adr ~ adults + children + meal + reservation_status
##
##                      Df Sum of Sq     RSS   AIC
## <none>                             3943113 15871
## - reservation_status  2     58068 4001181 15886
## - meal                3    196461 4139574 15950
## - adults              1    396008 4339121 16064
## - children            1    709532 4652645 16210
```

```
tidy(selected_model) %>%
  kable()
```

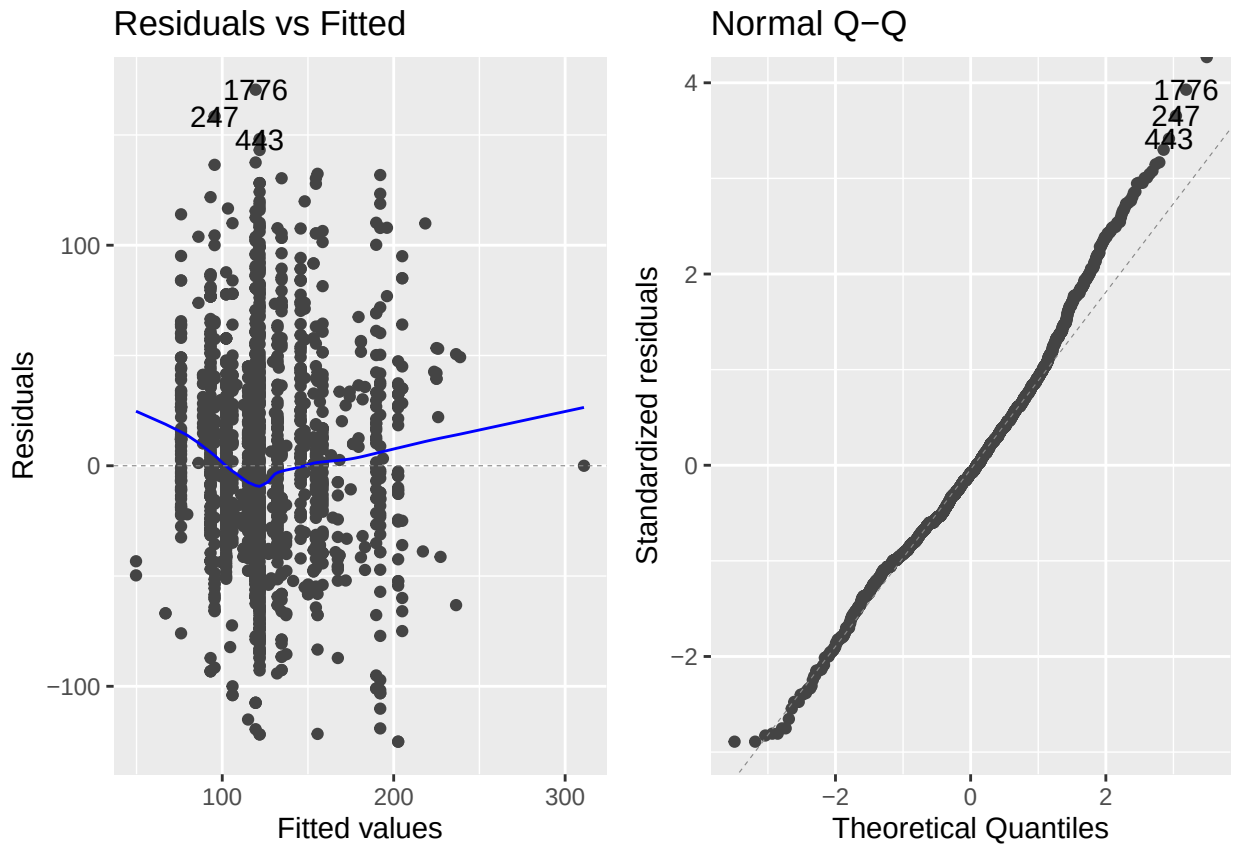
term	estimate	std.error	statistic	p.value
(Intercept)	80.018854	4.248222	18.8358463	0.0000000
adults	26.373831	1.820839	14.4844344	0.0000000
children	35.313159	1.821381	19.3881206	0.0000000
mealHB	33.743149	5.968921	5.6531405	0.0000000
mealSC	-17.818291	2.325765	-7.6612616	0.0000000
mealUndefined	107.607167	43.604964	2.4677734	0.0136752
reservation_statusCheck-Out	-12.666009	2.311291	-5.4800585	0.0000000
reservation_statusNo-Show	-2.626494	8.327674	-0.3153935	0.7524944

Modeling

```
model <- lm(adr ~ adults + children + babies + hotel + meal + reservation_status, data = hotels)
```

Conditions Check For Linear Model

```
# Residual + QQ Plots
autoplot(model, which = c(1,2))
```



• Linearity The residuals are randomly scattered in the plot of residuals vs fitted, so the linearity condition is satisfied. • Constant variance: The spread of the residuals is approximately equal in the plot of residuals vs. fitted as the fitted value increases. There are a few outliers, but overall the spread of the residuals is

equal. Therefore constant variance is satisfied. • Normality: The QQ-plot shows that the distribution of the residuals is skewed right, so this condition is not satisfied. However, the model is robust to deviations from normality for sample sizes greater than 30. Our sample size of 1447 is greater than 30, so we can proceed. • Independence: This condition is not met since there are multiple observations for most lemurs. These observations would be correlated

From this analysis, we conclude that the conditions are met for performing linear regression.

Results

```
tidy(model, conf.int = TRUE) %>%
kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	79.723	4.256	18.734	0.000	71.377	88.069
adults	26.277	1.825	14.398	0.000	22.698	29.856
children	35.169	1.825	19.274	0.000	31.590	38.747
babies	-22.637	17.815	-1.271	0.204	-57.574	12.300
hotelResort Hotel	2.353	2.391	0.984	0.325	-2.336	7.042
mealHB	33.779	5.982	5.647	0.000	22.048	45.511
mealSC	-17.205	2.409	-7.141	0.000	-21.930	-12.480
mealUndefined	106.033	43.632	2.430	0.015	20.466	191.600
reservation_statusCheck-Out	-12.809	2.323	-5.515	0.000	-17.364	-8.254
reservation_statusNo-Show	-2.662	8.327	-0.320	0.749	-18.991	13.668

$$\hat{AverageDailyRate} = 79.723 + 79.723(adults) + 35.169(children)$$
$$-22.637(babies) + 2.353(hotelResortHotel) + 33.779(mealHB) - 17.205(mealSc)$$
$$+106.033(mealUndefined) - 12.809(reservationstatusCheck-Out) - 2.662(reservationstatusNo-Show)$$

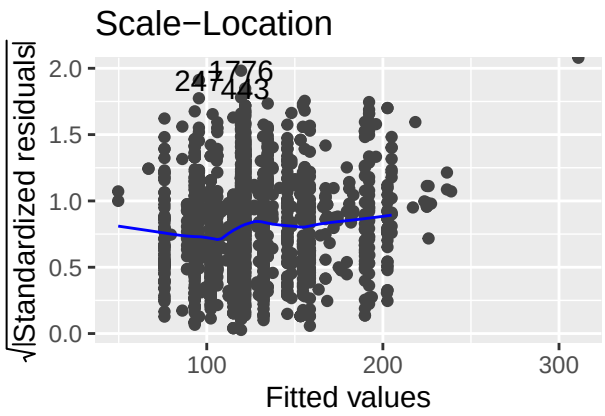
Diagnostics

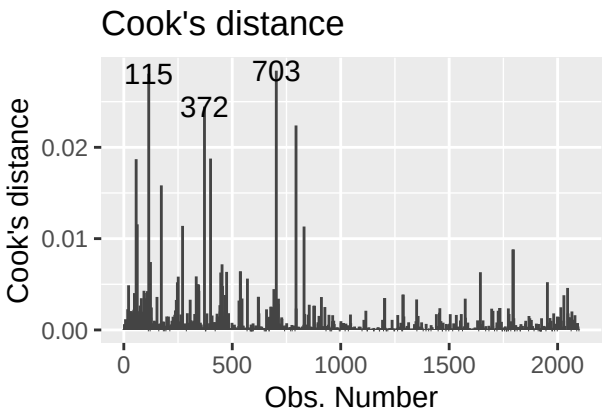
```
autoplot(model, which = c(3, 4, 5))
```

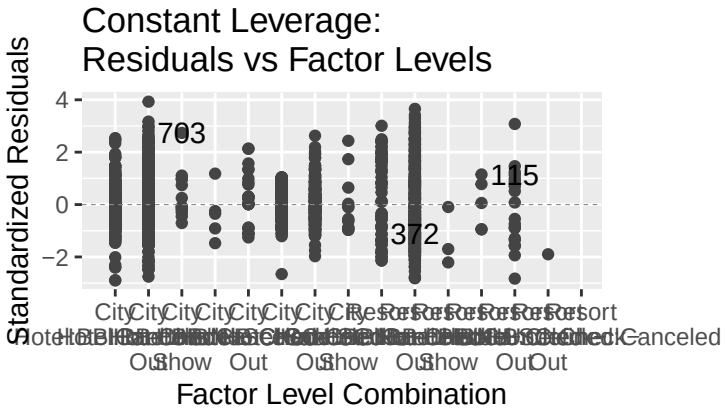
```
## Warning: Removed 14 row(s) containing missing values (geom_path).
```

```
## Warning: Removed 1 rows containing missing values (geom_segment).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```








```
#Leverage
lev_threshold <- 2 * 2 / nrow(hotels)
hotel_aug <- augment(model)

hotel_aug %>%
filter(.hat > lev_threshold) %>% nrow()
```

```
## [1] 1604
```

```
# High Magnitude Residual
hotel_aug %>%
filter(.std.resid < -3 | .std.resid > 3) %>% nrow()
```

```
## [1] 10
```

```
# Influential Point
hotel_aug %>%
filter(.cooksd > 1.0) %>% nrow()
```

```
## [1] 0
```

We use leverage and Cook's Distance to identify influential observations in our dataset, and use standardized residuals to identify outliers.

Leverage is a measure of the distance between an observation's values of the predictor variables and the average values of the predictor variables for the entire data set. We define a high leverage point as having a leverage greater than $ 2(p+1) /n $, where p is the number of predictors and n is the number of observations. We find X observations to be high leverage, and consider them to be potential influential points.

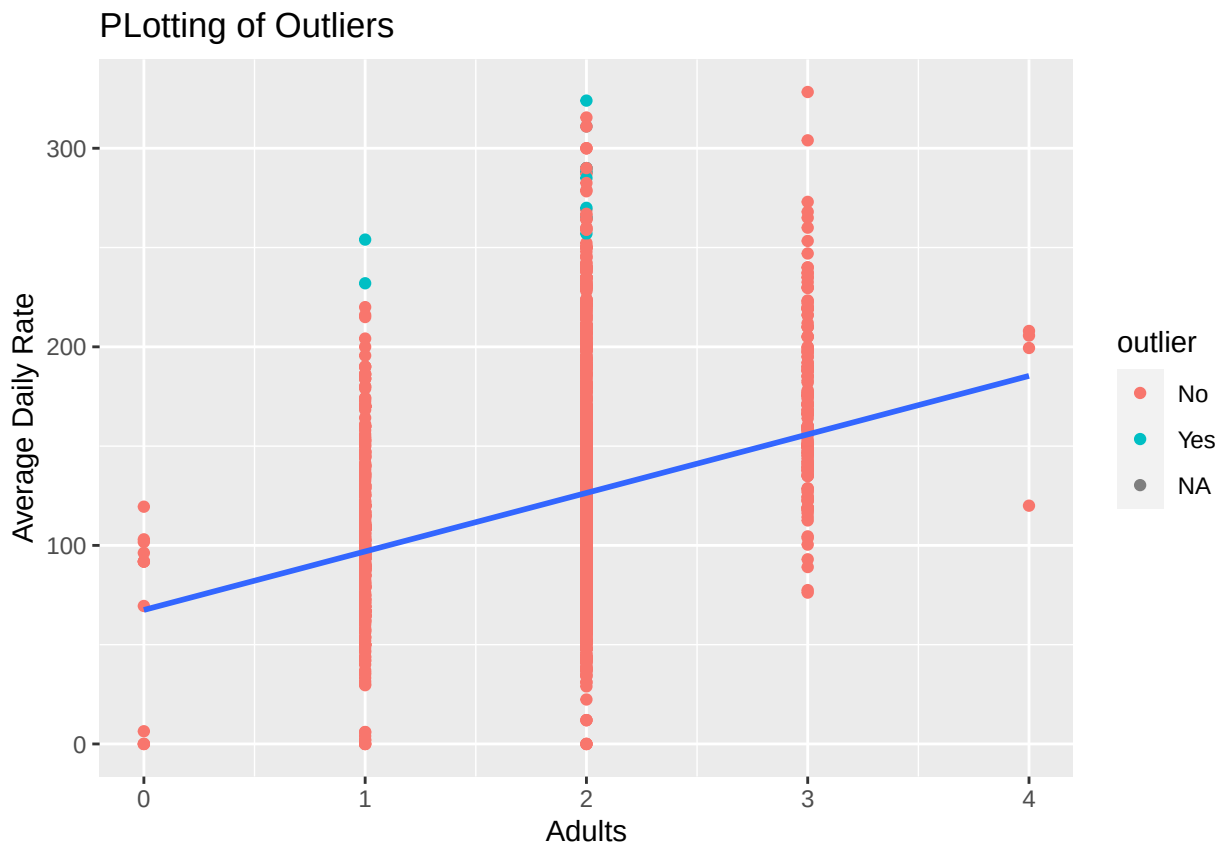
Cook's distance is a composite measure of an observation's leverage and standardized residual, and is used to identify influential points. An observation is considered an influential point if it's Cook's distance is greater than 1.0. We find X observations with a Cook's distance greater than 1.0, and conclude that they are influential points.

Standardized residuals can be used to identify outliers, as observations that have standardized residuals of large magnitude don't fit the pattern determined by the regression model. We identify outliers as observations with standardized residuals with a magnitude greater than or equal to 3. We find X observations to be outliers.

```
hotel_aug <- hotel_aug %>%
mutate(outlier = if_else(.std.resid < -3 | .std.resid > 3, "Yes", "No"))

ggplot(hotel_aug, aes(x = adults, y = adr)) + geom_point(aes(color = outlier)) +
geom_smooth(method = "lm", se = FALSE) + labs(title = "PLotting of Outliers",
      y = "Average Daily Rate",
      x = "Adults")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Model Assumptions, Limitations

Interpretations and Conclusions