Predicting the Cost of Hotel Booking in the United States

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

Nov 15th, 2021

Contents

1 References **19**

2 Appendix **19**

0.0.1 Introduction

Hotels are a critical component of the travel sector in the U.S., with an estimated 5.3 million guest rooms, and supporting 1 in 25 American jobs on average prior to 2020 based on this source. A 2018 report showed that approximately two thirds of Americans book their hotels directly through hotel websites, and a 2019 analysis showed large discrepancies in consumer travel cost based on how consumers chose to prioritize booking lodging compared to activities. We are interested in analyzing U.S. hotel bookings, with the goal of providing insight to consumers on the factors that affect their hotel cost.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 travelers to plan financially for future travel, or to potentially reduce cost. In this report, we are looking to use a variety of chosen models to understand the contributing factors to the average daily rate of a hotel room, as well as identify the strongest predictors.

0.0.2 Data

The source of the dataset is Tiny Tuesday. This data set was originally collected from an open hotel booking demand dataset (Antonio, de 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. 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.

0.0.3 Resesarch Question

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

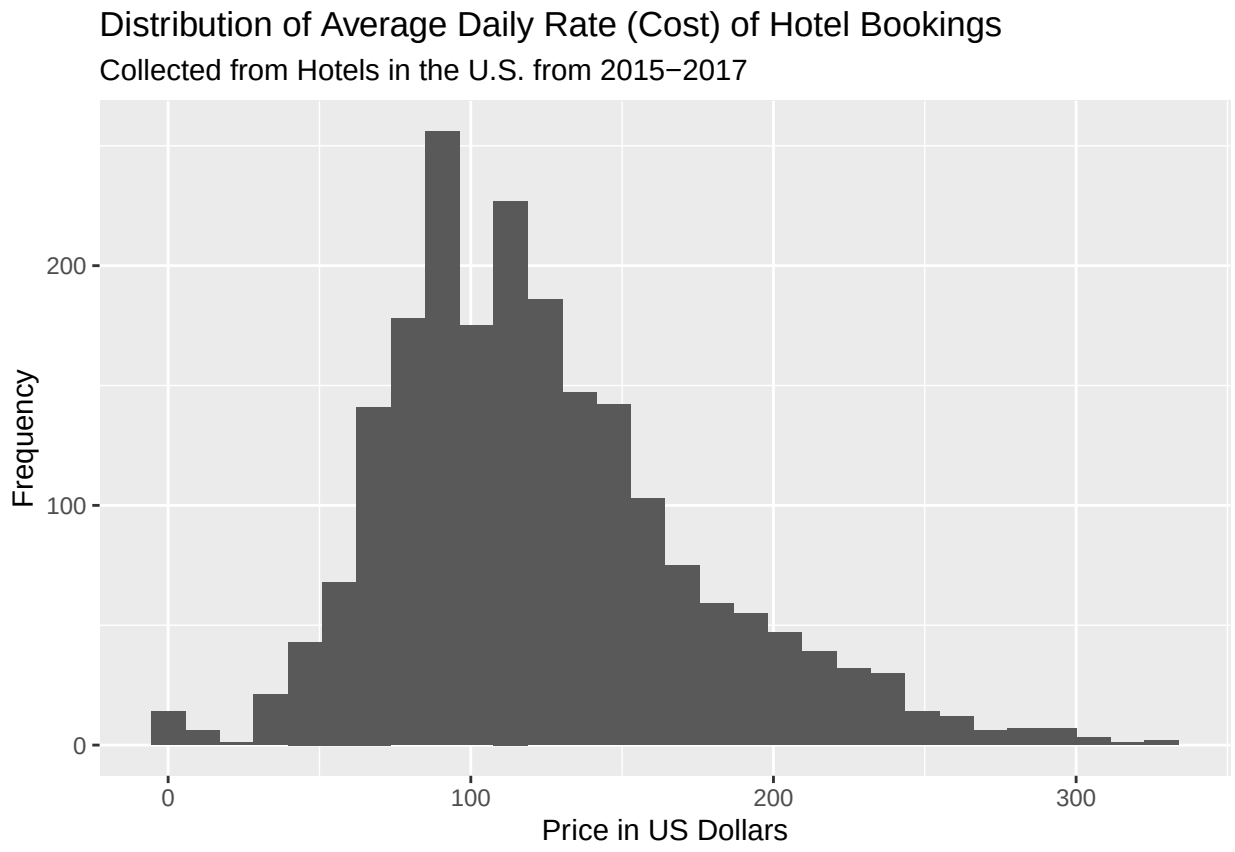
1) Evaluate the efficacy of this model to predict daily hotel room rates
2) Evaluate the statistical significance of each predictor, identifying the strongest contributing factors. We include the following variables in our analysis: hotel type, company, meal, number of adults/children/babies, the average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

0.0.4 Exploratory data analysis

#Appendix

```
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.

#Appendix

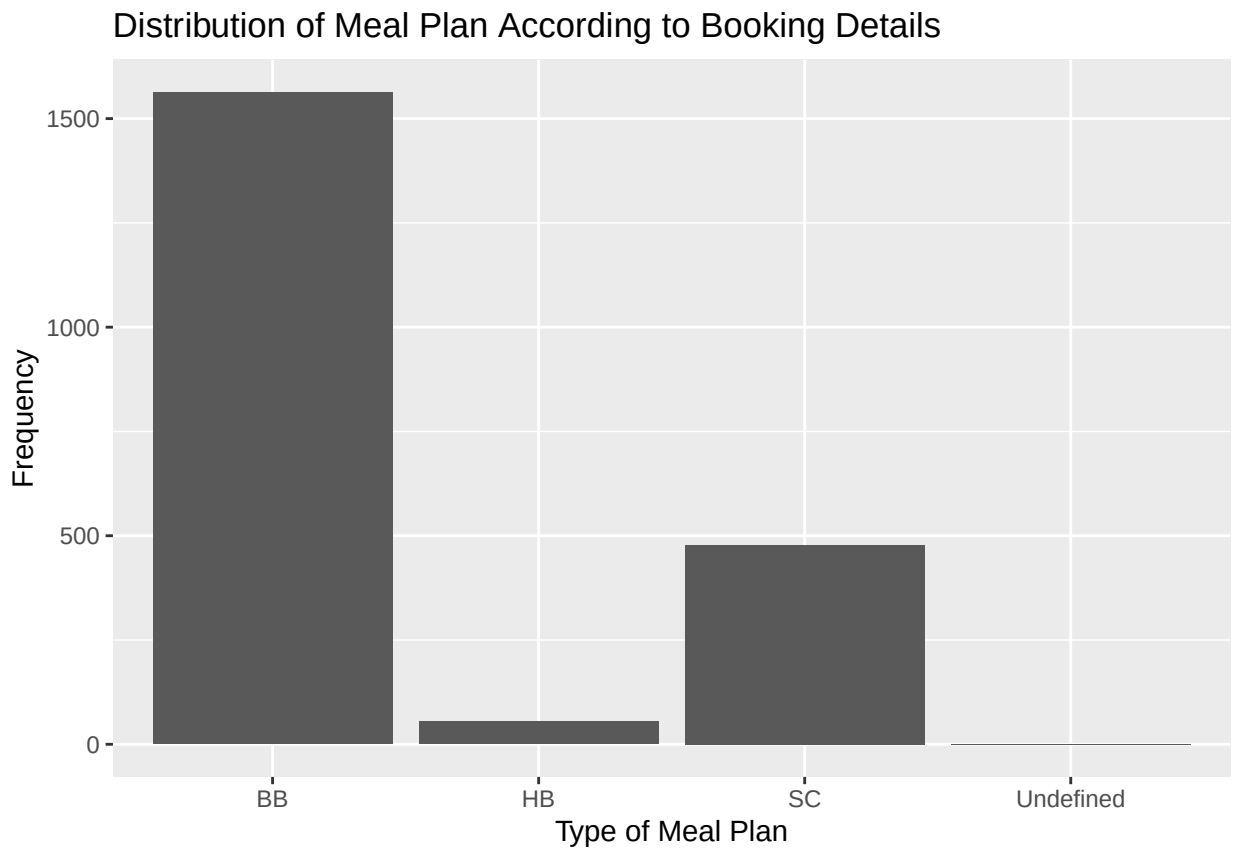
```
adults <- sum(hotels$adults)
children <- sum(hotels$children)
babies <- sum(hotels$babies)

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.

#Appendix

```
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")
```



#Appendix

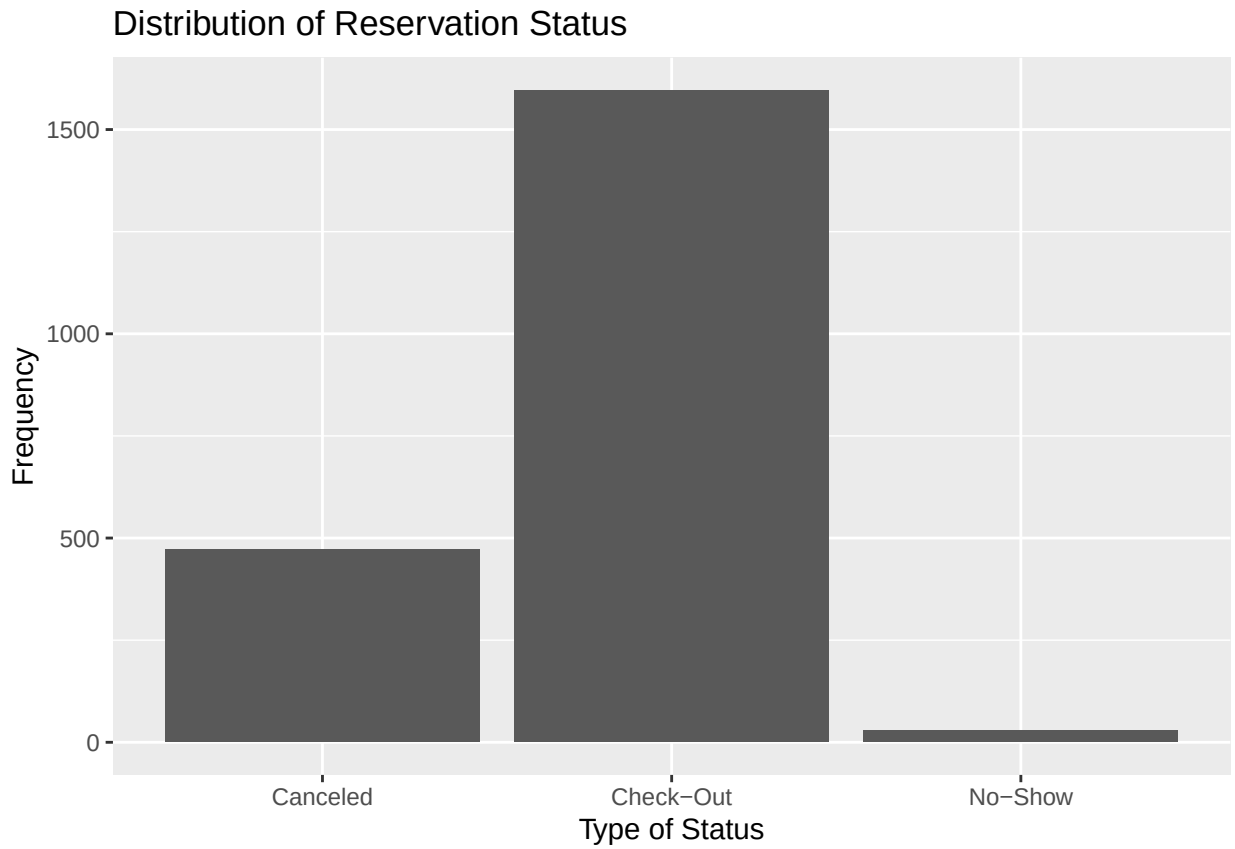
```
hotels %>%
  group_by(meal) %>%
  summarise(n = n(), mean = mean(adr), sd = sd(adr)) %>%
  kable(digits = 3)
```

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.

#Appendix

```
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")
```

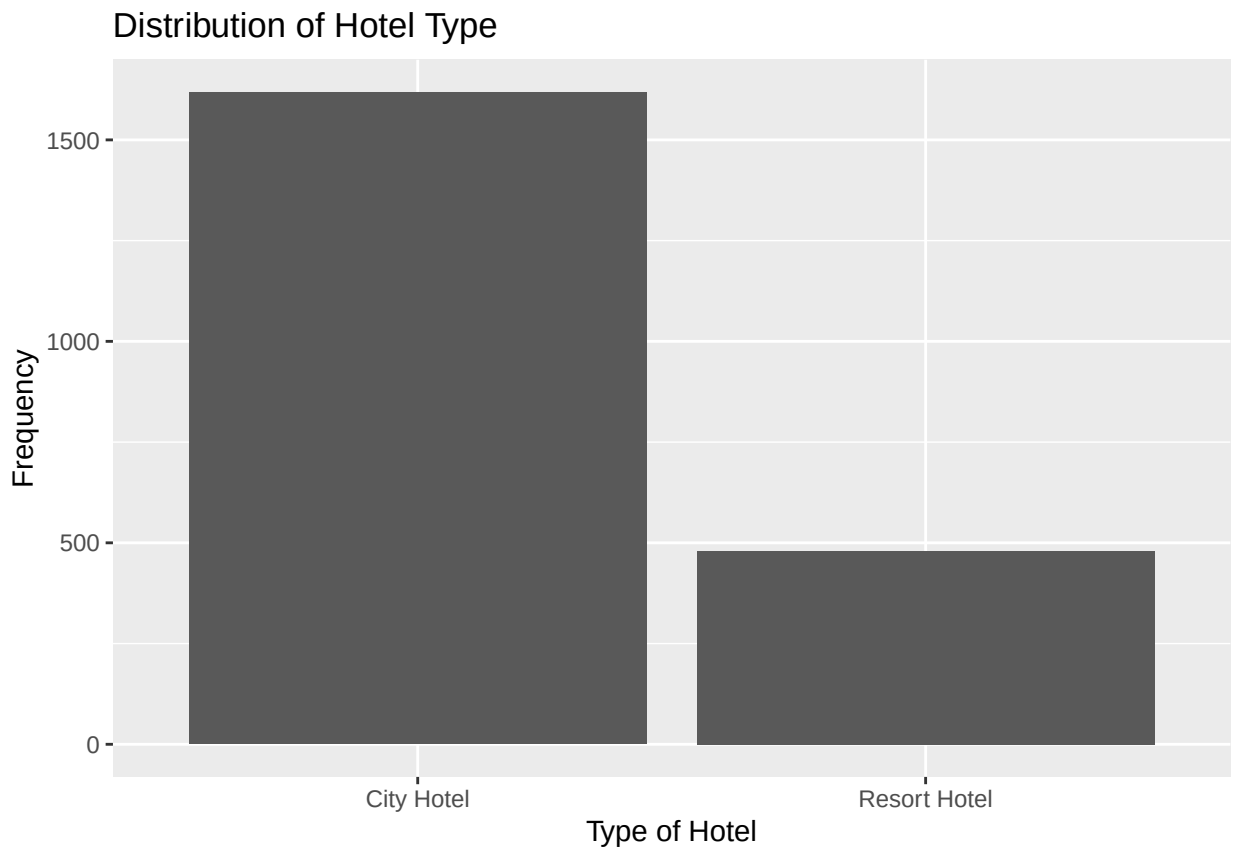
We cannot see any significant relationship between day in the month and the hotel price from this figure. However, we can see that summer is the busiest season for hotels and the price varies evenly. At the beginning of the year and the end of the year, hotels are not really booked as often.

```
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.

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




```
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
```

```
hotels <-hotels%>%
  mutate(children_present = case_when
         (hotels$adults > 0 & hotels$children == 0 & hotels$babies == 0 ~ 0,
          hotels$children > 0 | hotels$babies > 0 ~ 1),
         children_present = factor(children_present,
                                   levels = c(0,1)))

adult_adr <- hotels[hotels$children_present == 0,]$adr
kids_adr <- hotels[hotels$children_present == 1,]$adr

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

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

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.

0.1 Methodology

0.1.1 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.

```
if (hotels$arrival_date_week_number >=36 & hotels$arrival_date_week_number<=39)
{hotels$arrival_date_month = "September"}
```

```
## Warning in if (hotels$arrival_date_week_number >= 36 &
## hotels$arrival_date_week_number <= : the condition has length > 1 and only the
## first element will be used
```

```
if (hotels$arrival_date_week_number >=1 & hotels$arrival_date_week_number<=4)
{hotels$arrival_date_month = "January"}
```

```
## Warning in if (hotels$arrival_date_week_number >= 1 &
## hotels$arrival_date_week_number <= : the condition has length > 1 and only the
## first element will be used
```

```
hotels <- na.omit(hotels)
hotels <- subset(hotels, select = -country)
hotels <- subset(hotels, select = -reserved_room_type)
hotels <- subset(hotels, select = -assigned_room_type)
hotels <- subset(hotels, select = -agent)
hotels <- subset(hotels, select = -company)
hotels <- subset(hotels, select = -reservation_status_date)
hotels <- subset(hotels, select = -previous_cancellations)
hotels <- subset(hotels, select = -reservation_status)
hotels <- subset(hotels, select = -arrival_date_week_number)



hotels <- hotels%>%
  mutate(arrival_date_month = factor(arrival_date_month,
  levels = c("January", "February", "March", "April",
             "May", "June", "July", "August",
             "September", "October", "November",
             "December")),
           arrival_date_year = factor(arrival_date_year,
                                      levels = c(2015, 2016, 2017)),
```

```
          meal = case_when(meal == "SC" | meal == "Undefined" ~ "None",
                           meal == "BB" ~ "BB",
                           meal == "HB" ~ "HB"),
          meal = factor(meal, levels = c("None", "BB", "HB")))
```

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.

```
set.seed(12121)
hotel_split <- initial_split(hotels, prop = 0.8) #80% in the training set
hotel_train <- training(hotel_split)
hotel_test <- testing(hotel_split)
hotel_test <- hotel_test[!(hotel_test$meal=='Undefined'),]
rownames(hotel_train) <- NULL
rownames(hotel_test) <- NULL
nrow(hotels)
```

```
## [1] 2093
```

```
nrow(hotel_test)
```

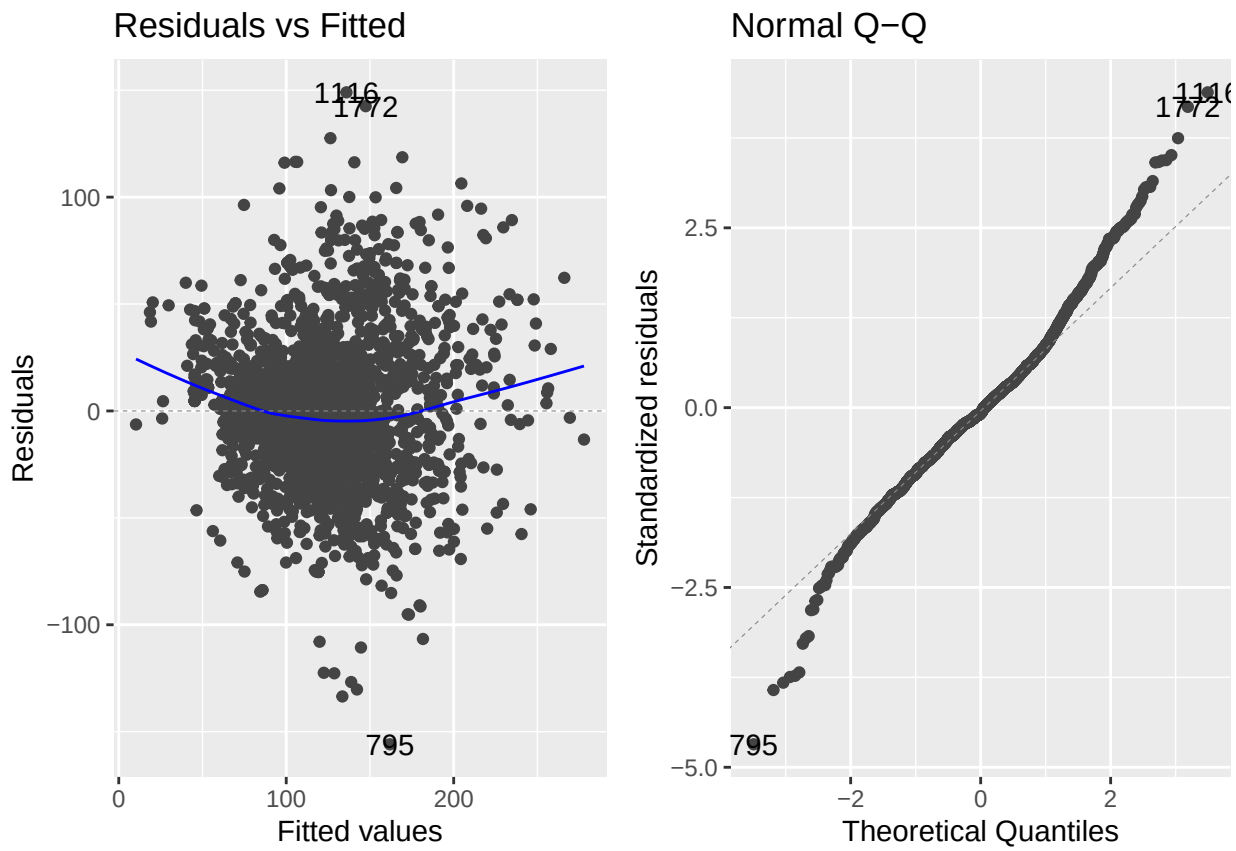
```
## [1] 419
```

```
full_model <- lm(adr ~ ., data = hotels)
#tidy(full_model) %>%
#  kable(digits = 3)
```

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

0.1.2 Conditions Check For Linear Model

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



To check that a linear model is applicable to this dataset, we check the following conditions: Linearity, Constant Variance, Normality, and Independence.

To check for Linearity, we look at the residuals vs fitted plot. We see that the residuals are randomly scattered, which indicates that the linearity condition is met.

To check for constant variance, we look at the spread of the residuals in the residuals vs fitted plot. We see that the spread of the residuals is approximately equal as the fitted value increases, indicating that the constant variance condition is satisfied.

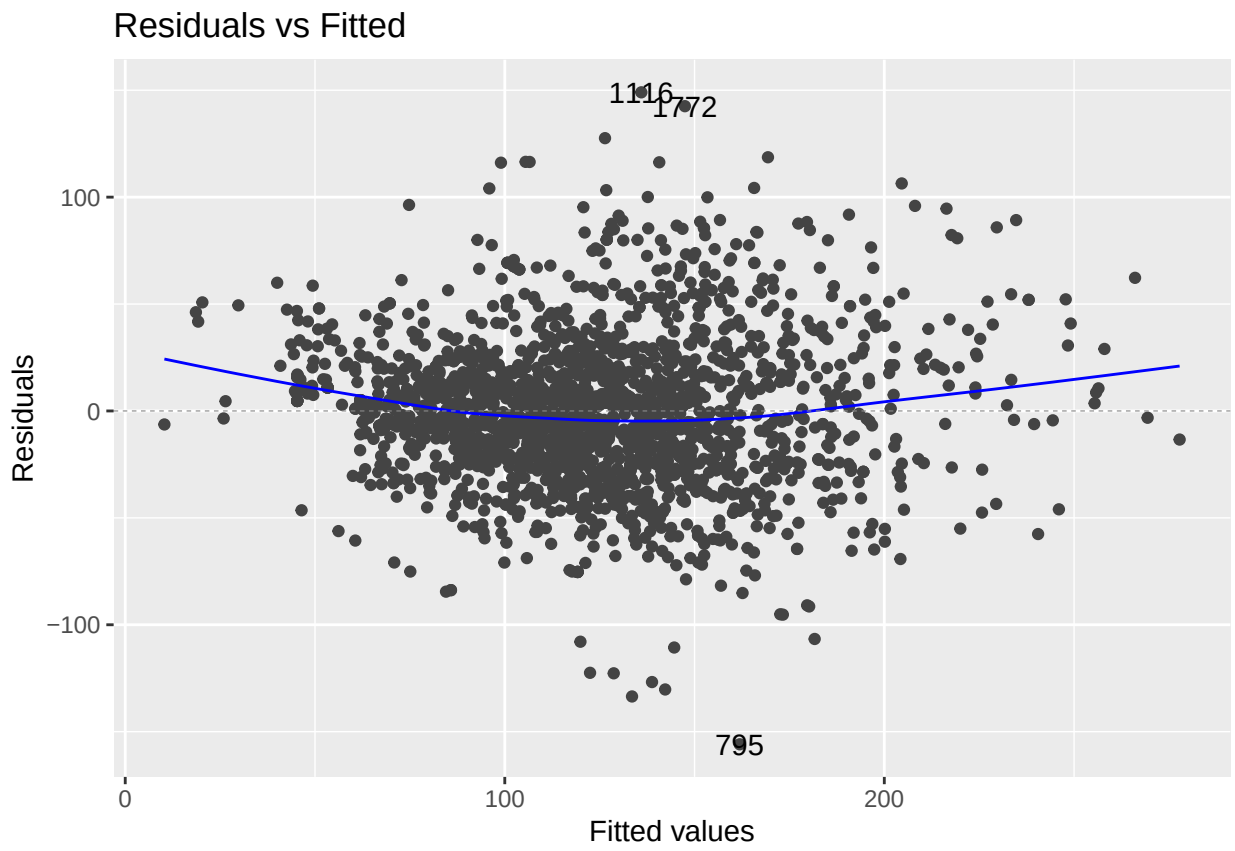
To check for normality, we look at the QQ-plot for a linear relationship. In the plot, we see a mostly linear relationship, and additionally recognize that our model is robust to deviations because our number of samples is much larger than 30. Therefore, we conclude that the normality condition is met.

To check for independence, we look at the study design, and find that there is no reason to believe that our samples are not independent from the survey, such that individual hotel bookings can be assumed to be independent.

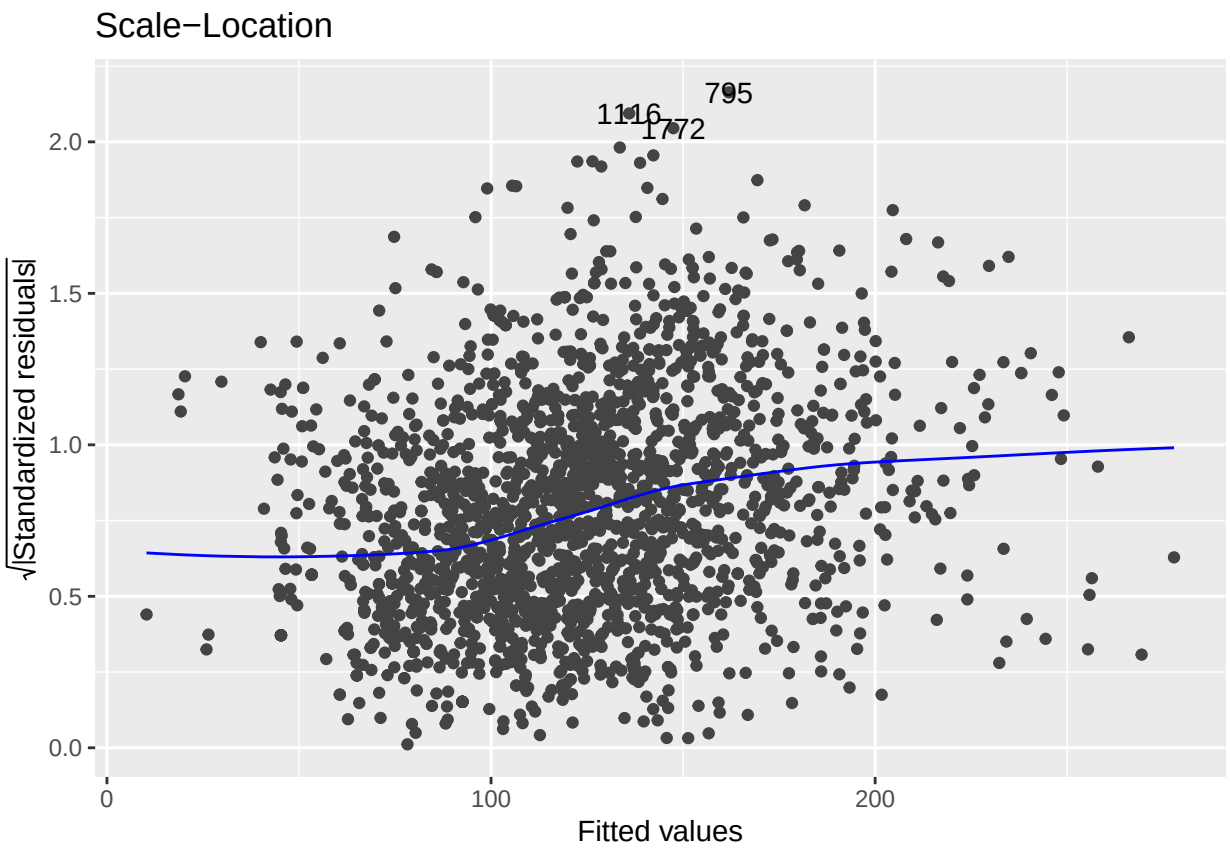
From this analysis, we conclude that the conditions are met for performing linear regression.

0.1.3 Diagnostics

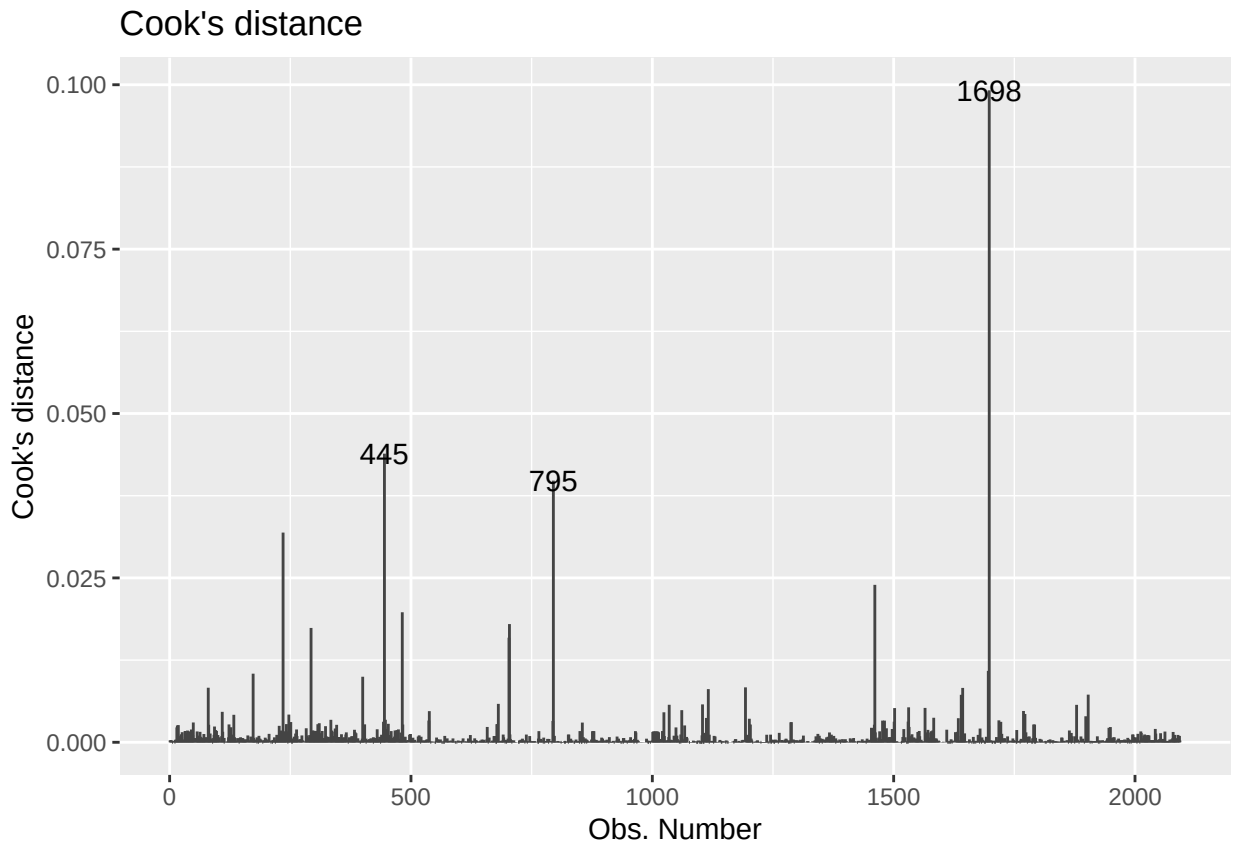
```
autoplot(selected_model, which = 1, ncol =1)
```




```
autoplot(selected_model, which = 3, ncol =1)
```




```
autoplot(selected_model, which = 4, ncol =1)
```




```
#Leverage

lev_threshold <- 2 * (16+1) / nrow(hotels)
hotel_aug <- augment(selected_model)

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

```
## [1] 448
```

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

```
## [1] 22
```

```
# Influential Point
hotel_aug %>%
filter(.cooksd > 0.5) %>%
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 448 observations to be high leverage, and consider them to be potential influential points.

Standardized residuals can be used to identify potential outliers, as observations that have standardized residuals of large magnitude don't fit the pattern determined by the regression model. We identify potential outliers as observations with standardized residuals with a magnitude greater than or equal to 3. We find 22 observations to be potential outliers.

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 a moderately influential point if it's Cook's distance is greater than 0.5. We find 1 observation with a Cook's distance greater than 0.5. Therefore, we can conclude that there is 1 moderately influential point.

After calculating the model diagnostics, we have found that on observation can be classified as an outlier. However, due to the fact that this is a legitimate observation of our dataset, we have chosen to keep this point in our model.

0.2 Results

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

```
model_interact <- lm(adr ~ hotel + is_canceled + lead_time + arrival_date_year + arrival_date_month + a

tidy(model_interact, conf.int =.95) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-11.392	9.667	-1.178	0.239	-30.350	7.567
hotelResort Hotel	-5.430	1.941	-2.797	0.005	-9.237	-1.623
is_canceled	9.622	1.915	5.024	0.000	5.866	13.377
lead_time	-0.177	0.010	-17.600	0.000	-0.197	-0.158
arrival_date_year2016	15.462	3.042	5.083	0.000	9.496	21.427
arrival_date_year2017	34.581	3.448	10.030	0.000	27.820	41.343
arrival_date_monthFebruary	0.906	10.663	0.085	0.932	-20.004	21.816
arrival_date_monthMarch	3.907	8.969	0.436	0.663	-13.683	21.497
arrival_date_monthApril	30.603	8.476	3.611	0.000	13.980	47.226
arrival_date_monthMay	52.286	7.725	6.768	0.000	37.136	67.436
arrival_date_monthJune	48.332	7.552	6.400	0.000	33.522	63.142
arrival_date_monthJuly	46.199	7.548	6.121	0.000	31.397	61.002
arrival_date_monthAugust	77.880	7.823	9.956	0.000	62.539	93.221
arrival_date_monthSeptember	51.447	8.161	6.304	0.000	35.442	67.453
arrival_date_monthOctober	62.318	8.492	7.339	0.000	45.664	78.971
arrival_date_monthNovember	48.329	10.033	4.817	0.000	28.654	68.005
arrival_date_monthDecember	-1.419	10.900	-0.130	0.896	-22.795	19.957
arrival_date_day_of_month	-0.562	0.446	-1.261	0.208	-1.435	0.312
stays_in_week_nights	1.513	0.521	2.903	0.004	0.491	2.536
adults	21.456	1.537	13.959	0.000	18.441	24.470
children	32.022	1.451	22.065	0.000	29.176	34.868
mealBB	21.739	1.994	10.904	0.000	17.830	25.649
mealHB	50.500	5.092	9.917	0.000	40.513	60.486
distribution_channelDirect	17.531	6.195	2.830	0.005	5.383	29.679
distribution_channelGDS	30.240	9.591	3.153	0.002	11.431	49.050
distribution_channelTA/TO	5.200	6.042	0.861	0.390	-6.650	17.050

term	estimate	std.error	statistic	p.value	conf.low	conf.high
is_repeated_guest	-53.217	9.627	-5.528	0.000	-72.097	-34.338
previous_bookings_not_canceled	45.803	11.550	3.966	0.000	23.152	68.454
days_in_waiting_list	-0.343	0.077	-4.436	0.000	-0.495	-0.192
total_of_special_requests	4.951	0.952	5.202	0.000	3.085	6.818
arrival_date_monthFebruary:arrival_date_day_of_month	0.088	0.671	0.131	0.896	-1.229	1.405
arrival_date_monthMarch:arrival_date_day_of_month	1.410	0.575	2.451	0.014	0.282	2.539
arrival_date_monthApril:arrival_date_day_of_month	0.965	0.533	1.809	0.071	-0.081	2.011
arrival_date_monthMay:arrival_date_day_of_month	0.464	0.505	0.919	0.358	-0.526	1.455
arrival_date_monthJune:arrival_date_day_of_month	0.577	0.500	1.154	0.249	-0.403	1.558
arrival_date_monthJuly:arrival_date_day_of_month	1.712	0.500	3.423	0.001	0.731	2.692
arrival_date_monthAugust:arrival_date_day_of_month	0.621	0.504	1.233	0.218	-0.367	1.610
arrival_date_monthSeptember:arrival_date_day_of_month	1.309	0.526	2.489	0.013	0.278	2.340
arrival_date_monthOctober:arrival_date_day_of_month	-0.583	0.569	-1.024	0.306	-1.699	0.534
arrival_date_monthNovember:arrival_date_day_of_month	-0.691	0.740	-0.934	0.350	-2.142	0.760
arrival_date_monthDecember:arrival_date_day_of_month	1.825	0.605	3.015	0.003	0.638	3.012

0.2.1 Interpretations and Conclusions

The model provides a number of useful predictors for determining how the average daily rate or price of a hotel is affected by its customization. The first aspect when choosing a hotel may be the type of hotel one would like to visit. In this model we looked at the difference between staying in a hotel located in a city/urban setting versus staying in a resort style hotel. According the model, the coefficient for a resort hotel is -5.430. This can be interpreted as if a guest decides to stay in a resort hotel versus a city hotel, they are expected to pay on average $5.43 less per day , holding all else constant. Because the p-value is very close to 0, we can also conclude that this is a significant statistic. Therefore, when deciding on a travel destination, it would be beneficial to consider the difference in price of a city or resort hotel.

Another important factor when planning a trip or booking a hotel is when you would like to travel or stay. The model covers a variety of predictors that involve the timing of a booking. For example, when considering which month to travel in, there is a possibility that depending on the month you book your stay, the average daily rate may be more expensive or less depending on the month. For example, the model identifies that the arrival date of a stay in August has a coefficient of 77.880. This can be interpreted as if a guest decides to book a stay during the month of August, they would be expected to pay on average $77.88 more per day, holding all else constant. Additionally, according to the model, if a guest stayed in a hotel with an arrival date in April, May, June, July, August, September, October, or November, they would be expected to pay on average a higher daily rate than staying in the month of January. This is proved by the fact that the p-value for each coefficient of these months is approximately 0, and the 95% confidence interval does not include 0 and has both thresholds being positive values. On the other hand, the months of February, March, and December all have fairly large p-values and have 0 fall in the 95% confidence interval. Therefore, we can say that these months are not statistically different from the month of January, meaning that there will not be a significant difference in the average daily rate whether a guest stays in January or February, March, or December.

Not only is the month of stay important, but the day within a month may also be an important factor in the price of a hotel stay. The model gives a coefficient of -0.562 for the arrival date day of the month, but has a p-value of 0.208. Because this p-value is larger than the $\alpha = 0.05$ level, we can say that the date of the month is not statistically significant. This is further proved by the 95% confidence interval (-1.435, 0.312), where 0 falls in the interval. Therefore, the arrival date in relation to the day of the month is not a useful predictor for the average daily rate of a hotel stay.

Additionally, when looking at the interaction between the day of the month and the month, we can see that maybe staying at a certain time within a certain month may have an effect in the cost. The months of March, July, September, and December all had p-values less than $\alpha = 0.05$ level, so we can conclude that they are

statistically significant. This means that the effect of Month of arrival date on the average daily rate is statistically different for July, September, and December.

Another attribute that may effect the daily price of the hotel is the number of guests and more specifically, the number of adults in the reservation. The model gives a coefficient of 21.456 and a p-value of 0 showing this coefficient is statistically significant. Therefore, we can conclude that for every additional adult added to the reservation, the average daily price of a hotel will increase on average $21.46, holding all else constant. From this, a pretty good estimate can be made about how much the daily rate will change with the number of adults under the reservation. This can similarly be applied to the number of children staying under a reservation. For every additional child added to the reservation, the average daily price of a hotel will increase on average $32.02, holding all else constant. We can be confident in this interpretation as the p-value is approximately 0 and 0 does not fall in the 95% confidence interval. Therefore, more children under a reservation will result in an increase of the average daily rate.

Lastly, we will discuss how a meal plan can possibly change the average daily price of a hotel stay. In the variable meal, the baseline is if a guest were to get no meal plan. The coefficient for Bed and Breakfast has a coefficient of 21.739 and a p-value of approximately 0. Therefore, we can say that if one were to choose the meal plan Bed and Breakfast, they would pay on average $21.74 more than if they did not buy a meal plan. Similarly for a Half Board meal plan (breakfast and one other meal), the guest is expected to pay on average $50.50 more per day for their hotel stay if they were to get the Half Board plan versus no meal plan. This interpretation is supported by the approximately 0 p-value for this coefficient.

In addition to discussing how the different predictors affect the average daily rate, one can also use the model to predict how much their hotel stay may cost in relation to these factors.

###PREDICTINGGGG

To determine the predictive ability of our model, we use the Root Mean Squared Error on both our training and test datasets below. We find a very similar error between our training and test sets, and find that on average, our predictions differ from the actual value by around $43.

```
#train_pred <- tibble(predicted = predict(model, hotel_train)) %>%
#                          bind_cols(hotel_train)
#train_pred %>%
# rmse(truth = adr, estimate = predicted)
#test_pred <- tibble(predicted = predict(model, hotel_test)) %>%
#                          bind_cols(hotel_test)
#test_pred %>%
#  rmse(truth = adr, estimate = predicted)
```

0.2.2 Model Assumptions, Limitations

Our model assumes that hotel bookings are independent from each other, which could potentially be broken in limited circumstances, such as competing hotel holiday promotions.

It's important to recognize that our model has several limitations. First, we recognize that the travel industry (including hotel booking) is incredibly vulnerable to tragic circumstances or disasters, which are not accounted for in our model. For example, we recognize that our model is not likely to characterize hotel bookings during COVID, given that our data is limited to the years 2015 to 2017. Additionally, we anticipate that the hotel room type (such as suite versus individual room) would be an important predictor of hotel room price, but is not included in our model. Our dataset included a variable for the room type, but it was mapped to a anonymous letter system (A,B,C..) to protect industry knowledge. We excluded this variable because it would be impossible to use in the real world without knowing what mapping was applied. In our exploratory analysis, we found that there were far more hotel bookings in cities than in resorts. Given this skew, our model may better characterize city bookings than resort bookings.

1 References

Antonio, Nuno, Ana de Almeida, and Luis Nunes. 2019. "Hotel Booking Demand Datasets." *Data in Brief* 22: 41–49. https://doi.org/https://doi.org/10.1016/j.dib.2018.11.126.

2 Appendix

2.1 Exploratory Data Analysis