Project (REPLACE WITH TITLE)

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

Nov 15th, 2021

```
library(tidyverse)
```

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

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

```
library(broom)
library(knitr)
library(ggfortify)
```

```
hotels <- read_csv("data/hotels_reduced.csv")
```

```
## Rows: 2097 Columns: 32

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (14): hotel, arrival_date_month, meal, country, market_segment, distribu...
## dbl (18): is_canceled, lead_time, arrival_date_year, arrival_date_week_numbe...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Introduction and Data

INSTRUCTIONS: This section includes an introduction to the project motivation, data, and research question. Describe the data and definitions of key variables. It should also include some exploratory data analysis. All of the EDA won't fit in the paper, so focus on the EDA for the response variable and a few other interesting variables and relationships.

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 the booking of hotels may start to increase. Therefore, 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 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.

In this report, we are looking to use linear models to understand the contributing factors to hotel room price, as well as identify the strongest predictors.

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.

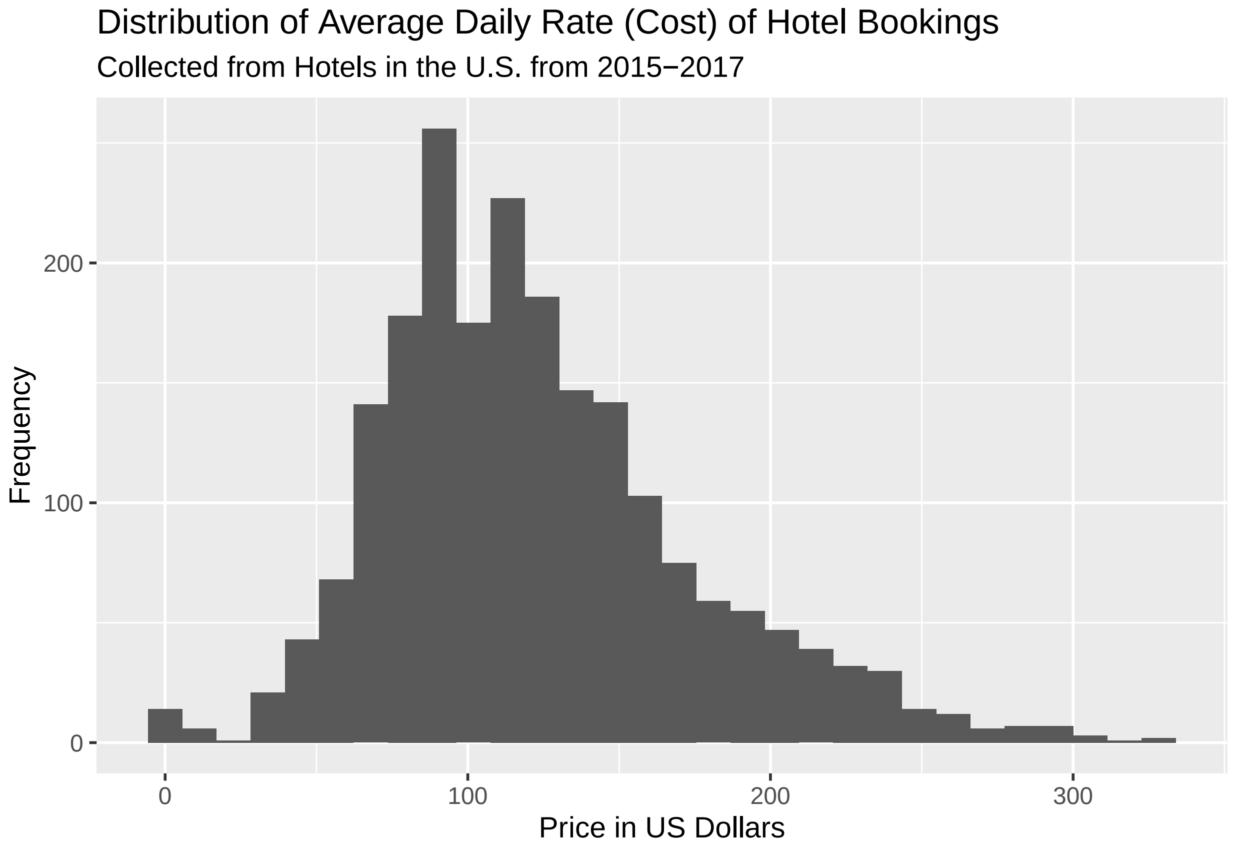
Resesarch Question

To better understand the contributing factors to the daily rate of hotel rooms in the U.S., we build a linear model to 1) evaluate the efficacy of this model to predict daily hotel room rates, and 2) evaluate the statistical significance of each predictor, identifying the strongest contributing factors. We include the following variables in our analysis: 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.

Exploratory data analysis

```r
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))
```

```
## # A tibble: 1 x 6
##    mean median    sd   min   max   iqr
##   <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  123.    115  51.6     0  328.  62.0
```

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.

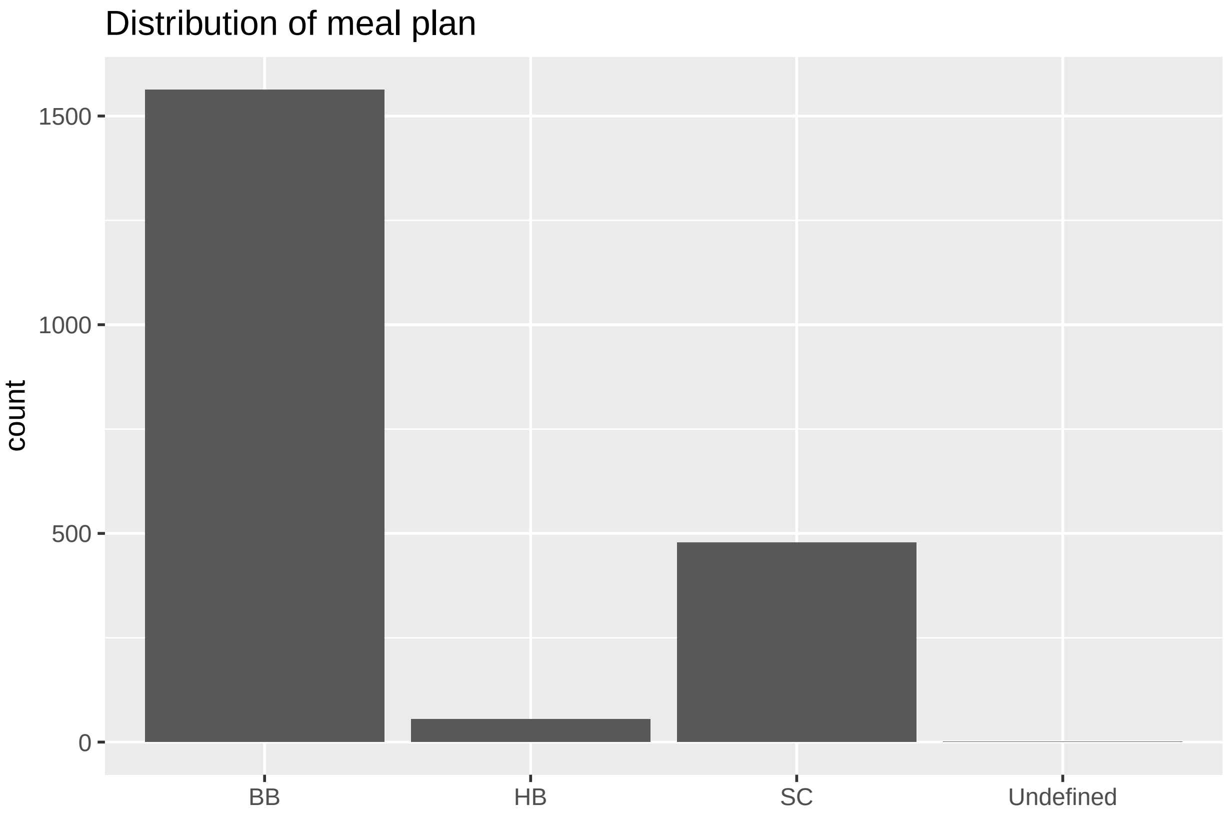
```
num_adults <- sum(hotels$adults)
num_children <- sum(hotels$children)
num_babies <- sum(hotels$babies)

tb <- table(num_adults, num_children, num_babies)
kable(tb)
```

num_adults	num_children	num_babies	Freq
3950	362	6	1

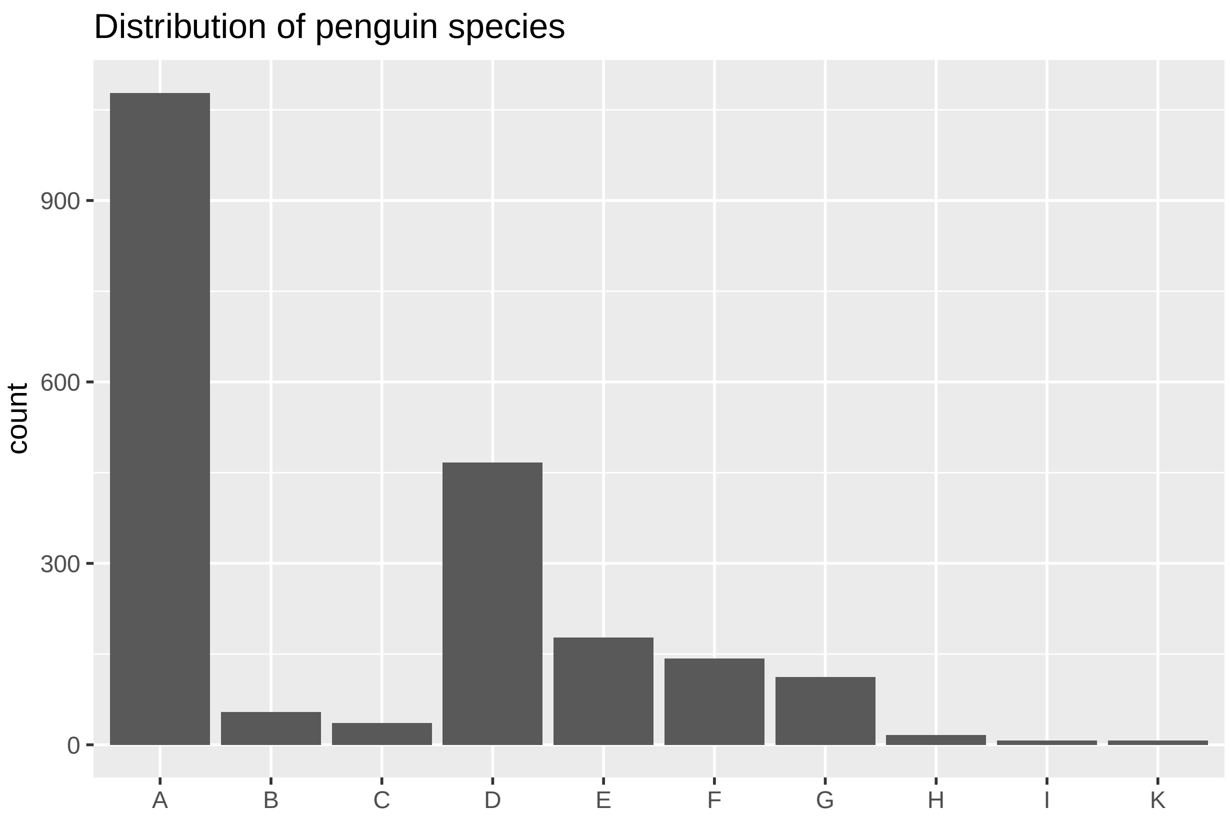
We can see that most of the users were adults and not many users come with babies.

```
ggplot(data = hotels, aes(x = meal)) +
geom_bar() +
labs(title = "Distribution of meal plan",
x = "")
```

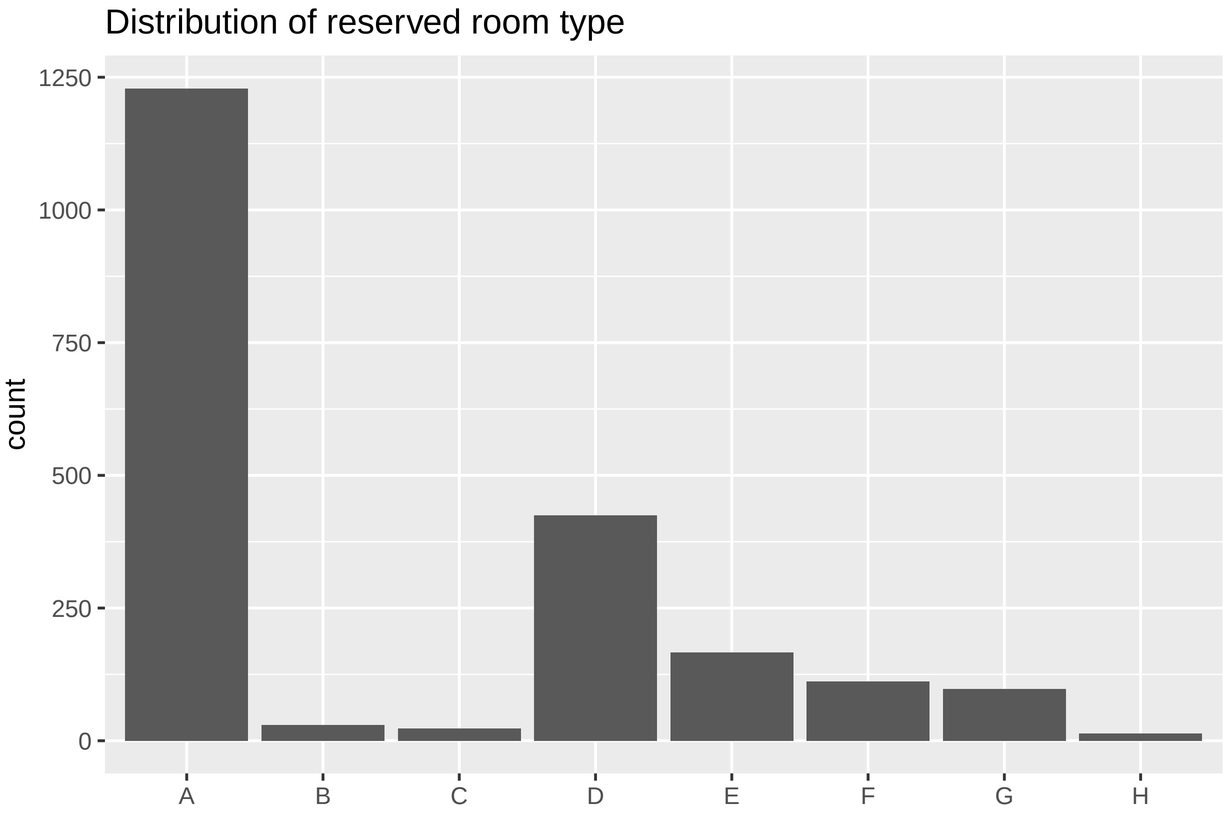


We can see that most of the people use bed and breakfast plan (BB) rather than half board or full board (HB and SC). Only few people chose no meal (undefined).

```
ggplot(data = hotels, aes(x = assigned_room_type)) +
geom_bar() +
labs(title = "Distribution of penguin species",
x = "")
```



```
ggplot(data = hotels, aes(x = reserved_room_type)) +
geom_bar() +
labs(title = "Distribution of reserved room type",
x = "")
```



## Methodology INSTRUCTIONS: This section includes a brief description of your modeling process. Explain the reasoning for the type of model you're fitting, predictor variables considered for the model including any interactions. Additionally, show how you arrived at the final model by describing the model selection process, any variable transformations (if needed), and any other relevant considerations that were part of the model fitting process.

Modeling Choices (Description)

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

As of now, we know we are going to use hotel (Resort Hotel or City Hotel), reserved_room_type, assigned_room_type, 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 possible predictor variables as we see fit.

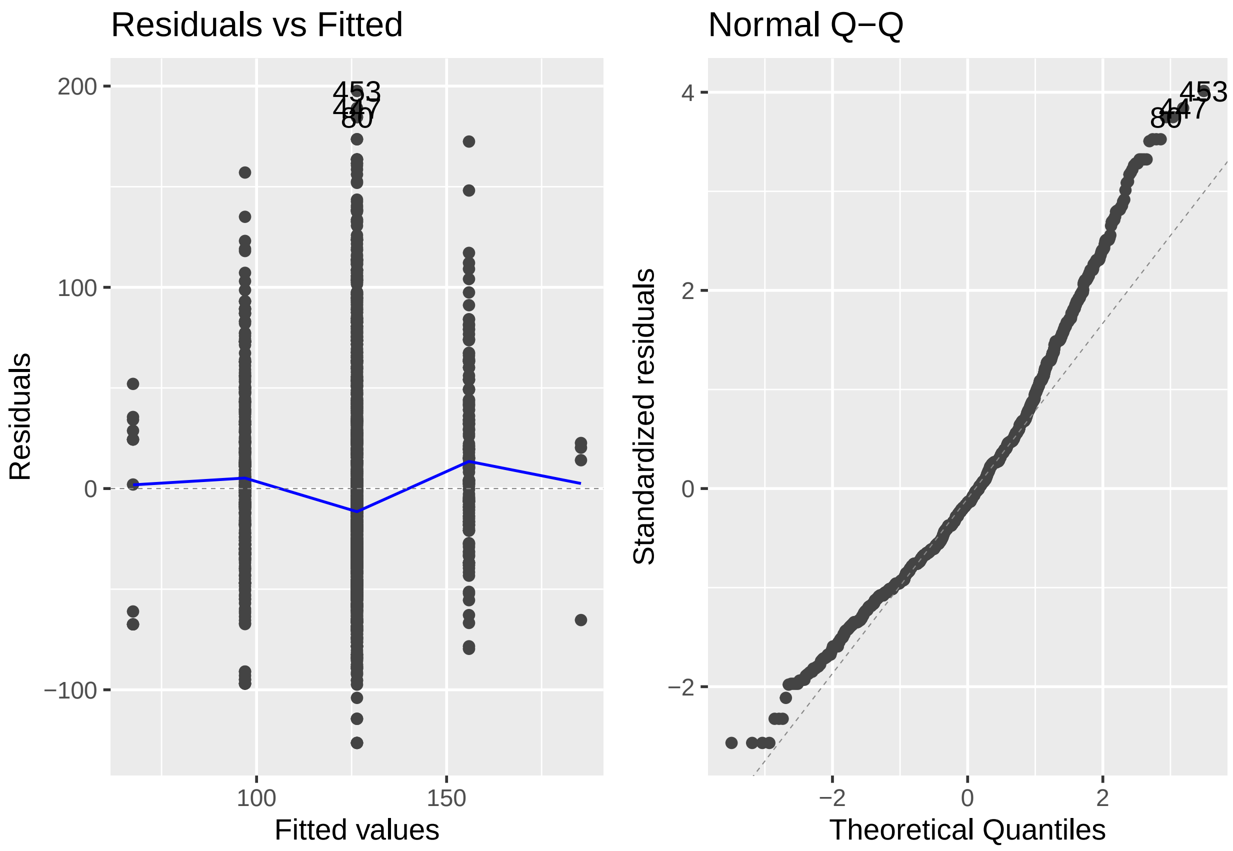
We plan to use multiple linear regression with a variety of combinations of interested predictor variables and their interaction. Based on whether conditions or fit or not or based on the visualization, we believe we may be utilizing logarithmic regression as well. We would also be interested in seeing 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.

Modeling

```
model <-lm(adr ~ adults, 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

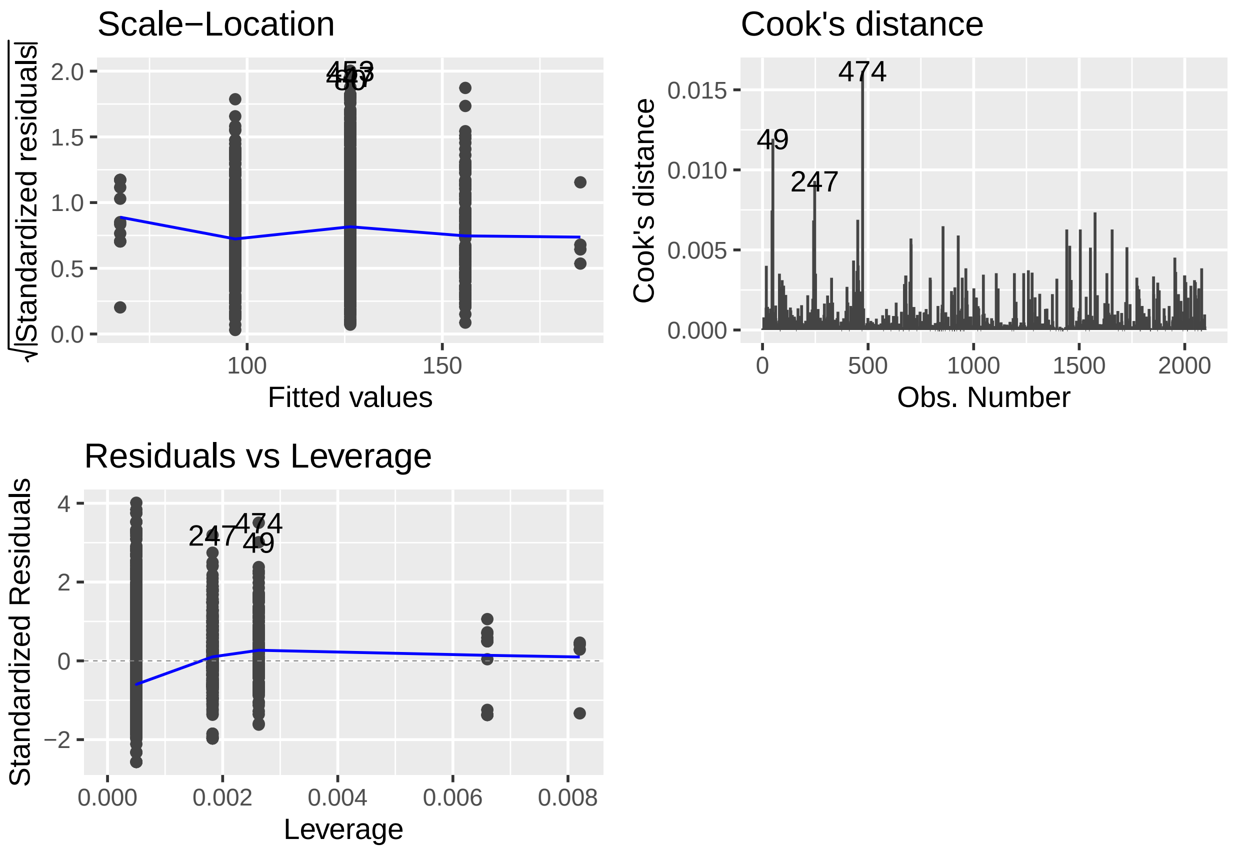
From this analysis, we conclude that the conditions are met for performing multiple linear regression.

Results

INSTRUCTIONS: In this section, you will output the final model and include a brief discussion of the model assumptions, diagnostics, and any relevant model fit statistics.This section also includes initial interpretations and conclusions drawn from the model.

Diagnostics

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



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

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

```
## [1] 174
```

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

```
## [1] 21
```

```
# 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.

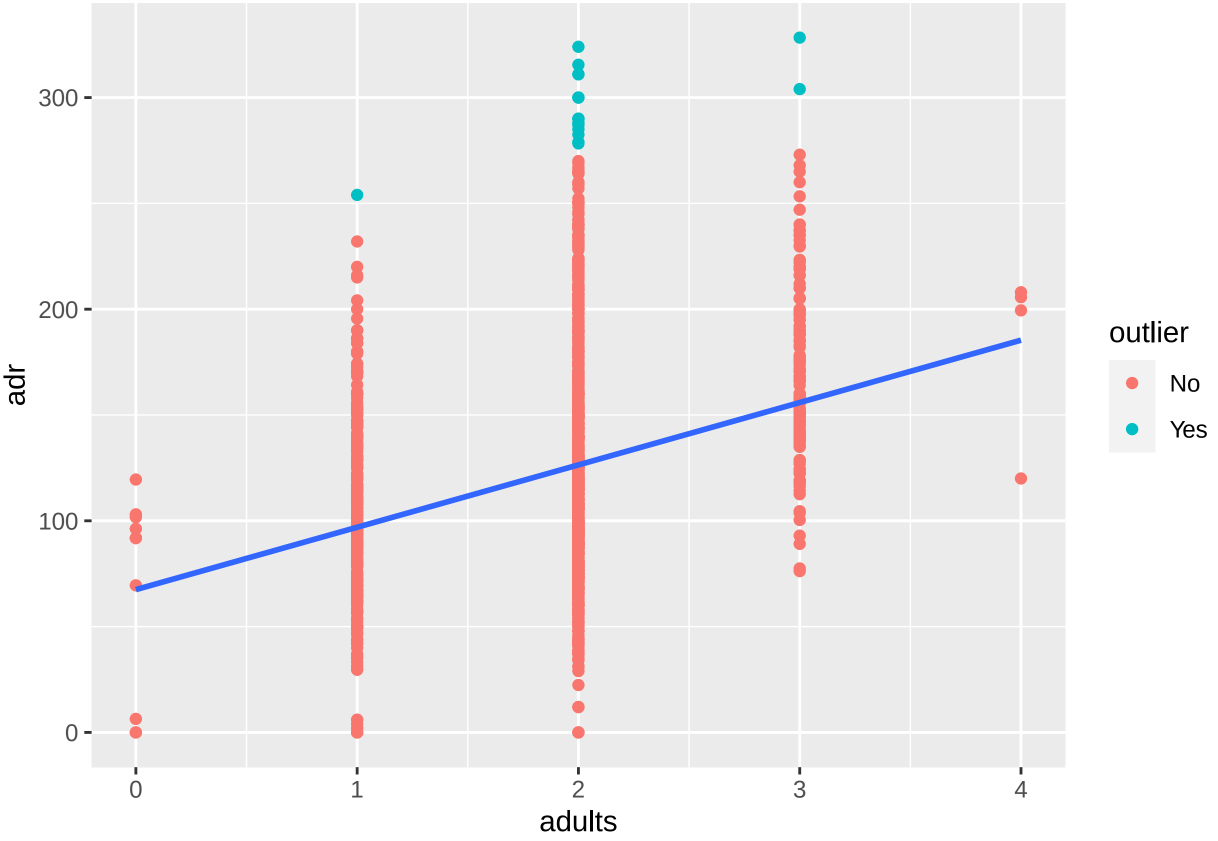
We hypothesize that these points are influential because _____.

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, and ________(remove them?).

```
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)
```

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



Model Assumptions, Limitations

Interpretations and Conclusions