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)
```

```
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.

Generally, we are looking to use linear models to understand the contributing factors to hotel room price, as well as identify the strongest predictors. We hypothesize that a model with predictors of 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, will be statistically significant predictor of hotel room price, and that the predictors will be significant except for company and number of adults/children/babies.

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.

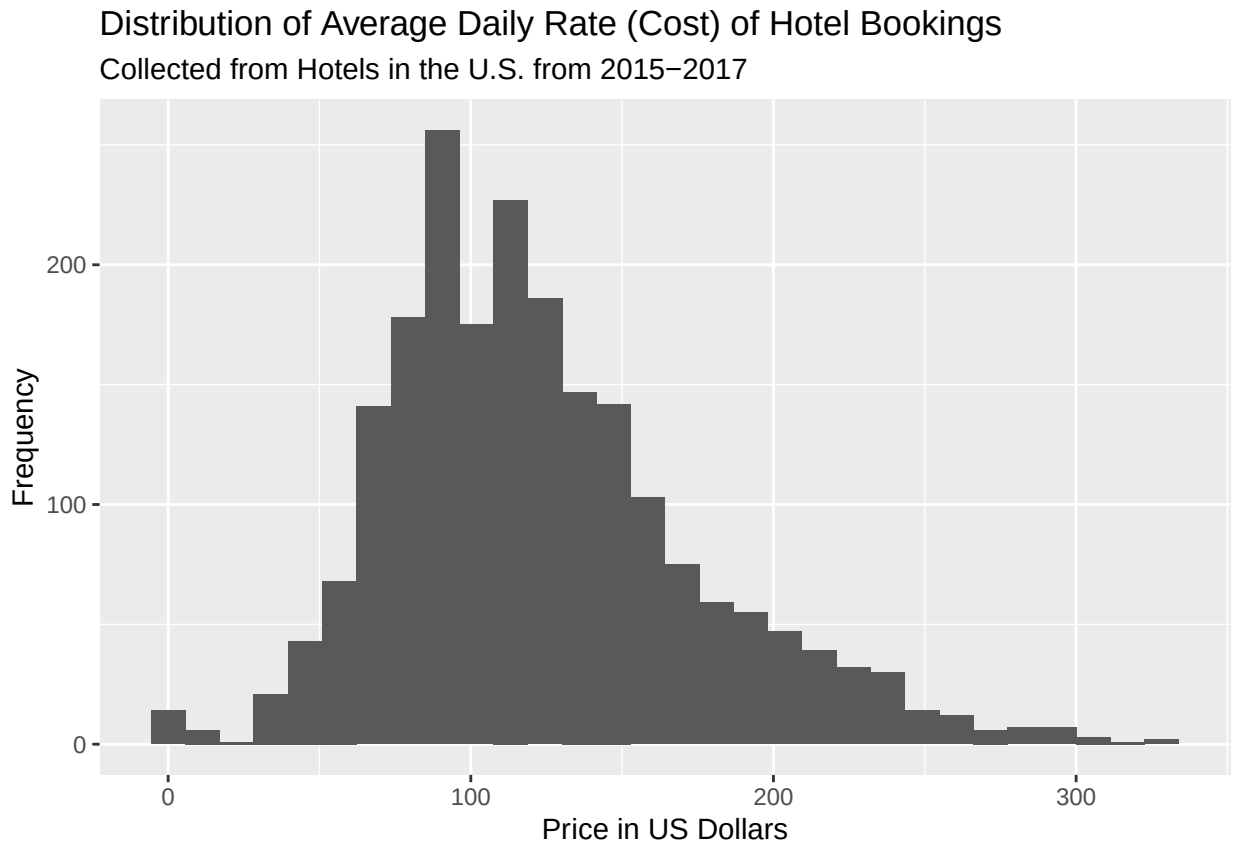
Resesarch Question

The main response variable we are interested in is the average daily rate or adr. The average daily rate can also be described as the daily cost of a hotel booking, and is calculated by dividing the sum of all lodging transactions by the total number of staying nights.

Exploratory data analysis

```
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 adr 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.

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.