Hotel Booking Demand: STA 210 Project Proposal

```
# load the data
hotels <- read_csv("data/hotel_bookings.csv")
```

Research Question

When talking about our group interests, it became clear that each of us has a passion for travel and other cultures. A lot of our co-curricular activities at Duke have been travel related, whether that be through Duke Engage, study abroad, or serving as the Head of Hospitality for Duke's club Business Oriented Women. We identified early on that we wanted to choose a data set that involved the travel industry and were interested in gaining insights that would lead us to be more informed explorers. As a result, we decided to investigate a hotel dataset, which has information about hotel booking demand. We would like to answer two research questions in order to investigate the trends in hotel booking throughout several years.

Our first question is: what factors of a hotel booking lead to the lowest average daily rate? In order to answer this question, we are going to build a multivariate regression model to predict the average daily price. The response variable for our regression model is `adr`, the average daily rate. The `adr` is calculated by dividing the sum of all lodging transactions by the total number of nights stayed. The explanatory variables we will use are `hotel`, `lead_time`, `arrival_date_month`, `stays_in_weekend`, `stays_in_week`, `adults + children + babies`, `meal`, and `distribution_chanel`. `Hotel` is a categorical variable that records whether the hotel is a city hotel or a resort hotel, and we hypothesize that the average daily rate would be less for city hotels, since resort hotels are generally more upscale and have more amenities. `Lead_time` is a quantitative variable that records the number of days between when the booking was made and the arrival date of the guests. We hypothesize that the greater the lead time, the lower the average daily rate, as booking in advance generally leads to the best rates. `Arrival_date_month` is a categorical variable that records the month of the arrival date, and we think that the arrival month might have an impact on the rate due to seasonal demand or major holidays. `Stays_in_weekend` is a quantitative variable that records the number of weekend nights the guests stayed at the hotel. We hypothesize that for more nights stayed on a weekend, the greater the average daily rate would be, since weekends are busy times for travel, as many people have weekends off. Likewise, `stays_in_week` is a quantitative variable that records the number of week nights the guests stayed at the hotel. We anticipate that bookings with a greater proportion of weekday nights than weekend nights will have a lower average daily rate. We plan to create a variable `stays_in_days` that combines the values of `stays_in_weekend` and `stays_in_week` to allow for a more straightforward analysis of length of stay. We will create the variable `total_guests` from `adults`, which records the number of adults, `children`, which records the number of children, and `babies`, which records the number of babies. Combining the three variables into one will allow for more straightforward analysis, and we predict that the greater the value is for `total_guests`, the greater the average daily rate will be. `Meal` is a categorical varaible that records the type of meal booked. We predict that more inclusive meal plans will have a higher average daily price assigned to them. Finally, the `distribution_channel` can either be direct, through a travel agent, or a tour operator. We predict that booking directly will lead to a lower average daily rate, as hotels prefer that clients book directly and reward them by doing so with lower rates.

Our second research question is: Can we predict whether a customter is going to cancel their booking? To answer this question, we will build a logistic regression model. The response variable for our logistic regression model is `is_canceled`. `is_canceled` Is a categorical variable that has values 1 and 0. 1 corresponds to a cancelled booking and 0 corresponds to non-cancelled bookings. The explantory variables for our logistic regression model are: `Lead_time`, since booking too far in advance might lead to unforseen conflicts, `children + babies`, as infants can get sick easily and their parents might need to stay home, `stays_in_weekend_nights` and `stays_in_week_nights`, since people might be more inclined to cancel trips on weekdays if they cannot get off of work, `is_repeated _guest`, since repeated guests might be more loyal to their hotel bookings, `previous_cancellations`, as people who are serial cancellers might be inclined to cancel again, `booking_changes`, since people who make a lot of changes to a booking might be on the fence about traveling, and finally, `deposit_type`, since "Non-refundable" trips are not likely to be cancelled.

Description of the Data

`hotels` contains hotel demand data. There are 119,390 observations, each representing a hotel booking. The bookings were retrieved from one city hotel in Lisbon, Portugal, and one resort hotel in Algarve, Portugal. Both datasets include bookings due to arrive between July 1, 2015, and August 31, 2017, including bookings that effectively arrived and bookings that were canceled.

The data set includes 32 variables of interest, including information about each booking, such as whether the booking was cancelled (`is_cancelled`), the customer's meal plan (`meal`), room type (`assigned_room_type`), when the booking was made (`arrival_date_year`, `arrival_date_month`, `arrival_date_day`), the length of the stay (`stay_weekend_nights`, `stay_week_days`), the number of adults, children, and/or babies (`adults`, `children`, `babies`), and the number of required parking spaces (`required_car_parking_spaces`), among others. A detailed description of all 32 variables can be found in the code book.

The data was originally from the article "Hotel Booking Demand Datasets," written by Nuno Antonio, Ana Almeida, and Luis Nunes for the journal *Data in Brief*, Volume 22, which was published in February 2019. The data set we will be using was cleaned by Thomas Mock and Antoine Bichat during the week of February 11th, 2020 for #TidyTuesday.

In the article "Hotel Booking Demand Datasets," the authors state that the data was extracted from the hotels' public Property Management System (PMS) databases' servers "by executing a TSQL query on SQL Server Studio Manager, the integrated environment tool for managing Microsoft SQL databases." The article assured that there was no missing data in the data sets constructed. Since this is real hotel data, all identifying data elements about the hotel or the customer were deleted.

While the primary table used to compile the data set was "Bookings," the researchers joined the "Bookings" table with other tables, including "Bookings change log," "Meals," "Distribution Channels," "Transactions," "Customer Profiles," "Nationalities," and "Market Segments," to get a more complete picture of the variables that affect bookings.

The documentation for the original data set can be found here: https://www.sciencedirect.com/science/article/pii/S2352340918315191

The Data

Here we will take a glimpse of the data:

```
glimpse(hotels)
```

```
## Observations: 119,390
## Variables: 32
## $ hotel                         <chr> "Resort Hotel", "Resort Hotel",...
## $ is_canceled                   <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1...
## $ lead_time                     <int> 342, 737, 7, 13, 14, 14, 0, 9, ...
## $ arrival_date_year             <int> 2015, 2015, 2015, 2015, 2015, 2...
## $ arrival_date_month            <chr> "July", "July", "July", "July",...
## $ arrival_date_week_number      <int> 27, 27, 27, 27, 27, 27, 27, 27,...
## $ arrival_date_day_of_month     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ stays_in_weekend_nights       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ stays_in_week_nights          <int> 0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 4...
## $ adults                        <int> 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2...
## $ children                      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ babies                        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ meal                          <chr> "BB", "BB", "BB", "BB", "BB", "...
## $ country                       <chr> "PRT", "PRT", "GBR", "GBR", "GB...
## $ market_segment                <chr> "Direct", "Direct", "Direct", "...
```

```
## $ distribution_channel           <chr> "Direct", "Direct", "Direct", "...
## $ is_repeated_guest              <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ previous_cancellations         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ previous_bookings_not_canceled <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ reserved_room_type             <chr> "C", "C", "A", "A", "A", "A", "...
## $ assigned_room_type             <chr> "C", "C", "C", "A", "A", "A", "...
## $ booking_changes                <int> 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ deposit_type                   <chr> "No Deposit", "No Deposit", "No...
## $ agent                          <chr> "NULL", "NULL", "NULL", "304", ...
## $ company                        <chr> "NULL", "NULL", "NULL", "NULL",...
## $ days_in_waiting_list           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ customer_type                  <chr> "Transient", "Transient", "Tran...
## $ adr                            <dbl> 0.00, 0.00, 75.00, 75.00, 98.00...
## $ required_car_parking_spaces    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ total_of_special_requests      <int> 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0...
## $ reservation_status             <chr> "Check-Out", "Check-Out", "Chec...
## $ reservation_status_date        <date> 2015-07-01, 2015-07-01, 2015-0...
```