1. Data Collection

The dataset being used is the [Hotel Booking Demand dataset from Kaggle](https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand).

It contains booking information for a city hotel and a resort hotel, including details like booking lead time, cancellation policies, customer types, and various other factors.

It was then loaded into Google Colab, a Jupyter Notebook service. The dataset was read into a DataFrame using pd.read\_csv(), and a copy of the original data was saved. Missing values were addressed using df.dropna() to simplify analysis. The dataset was further split into two separate dataframes for City Hotel and Resort Hotel bookings

.

1. Exploratory Data Analysis and Data Cleaning

Descriptive Statistics: df.describe() was used to get summary statistics like mean, standard deviation, minimum, and maximum values for numerical columns. This helps in identifying the range and distribution of values in the dataset. Some columns that were deemed unlikely to contribute significantly to cancellation prediction have been dropped, simplifying the dataset and preparing it for predictive modeling:

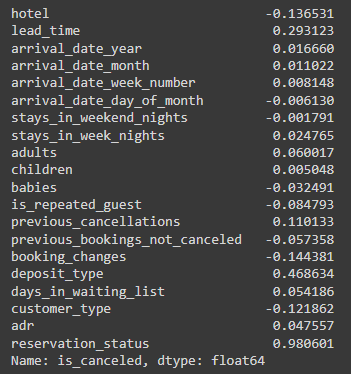
1. meal
2. country
3. market\_segment
4. Distribution\_channel
5. reserved\_room\_type
6. assigned\_room\_type
7. agent, company
8. required\_car\_parking\_spaces
9. total\_of\_special\_requests
10. reservation\_status\_date

Next we converted various categorical variables to numeric:

1. deposit\_type: Replaced 'No Deposit', 'Non Refund', and 'Refundable' with 0, 1, and 2
2. reservation\_status: Transformed 'Check-Out', 'Canceled', and 'No-Show' into numerical values 0, 1, and 2.
3. hotel: City Hotel was mapped to 0, and Resort Hotel was mapped to 1.
4. customer\_type: Various customer types such as Transient, Transient-Party, Contract, and Group were converted into 0, 1, 2, and 3.
5. Date Transformation: The arrival date (broken into year, month, and day) was combined into a single date\_booked column, which was then used to create a booked\_on column by subtracting the lead\_time.

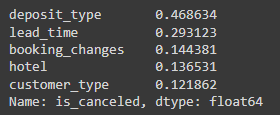
3. Potential predictor variables:

First we began by calculating the correlation coefficients of each variable with cancellation. Here, “reservation\_status” represents the last known status of the booking. The reason it doesn’t correlate perfectly with cancellation status is because some reservations are adjusted without being canceled, however the vast majority of adjustments are cancellations.

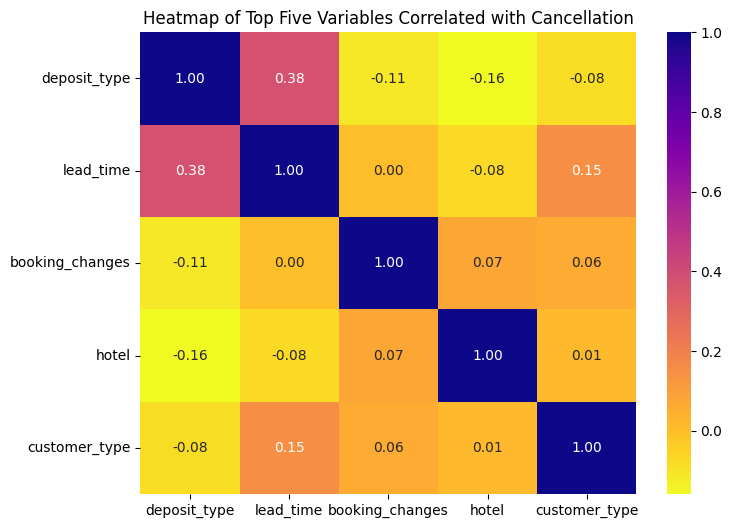


The top five correlators to cancellation were:

1. Deposit Type: Whether the booking required a deposit and what type of deposit was made.
2. Lead Time: The number of days between the booking and the actual arrival.
3. Booking Changes: adjustments such as rescheduling a booking
4. Hotel Type: Whether the booking was made at a city hotel or resort hotel.
5. Customer Type: Different customer types may have different booking behaviors.



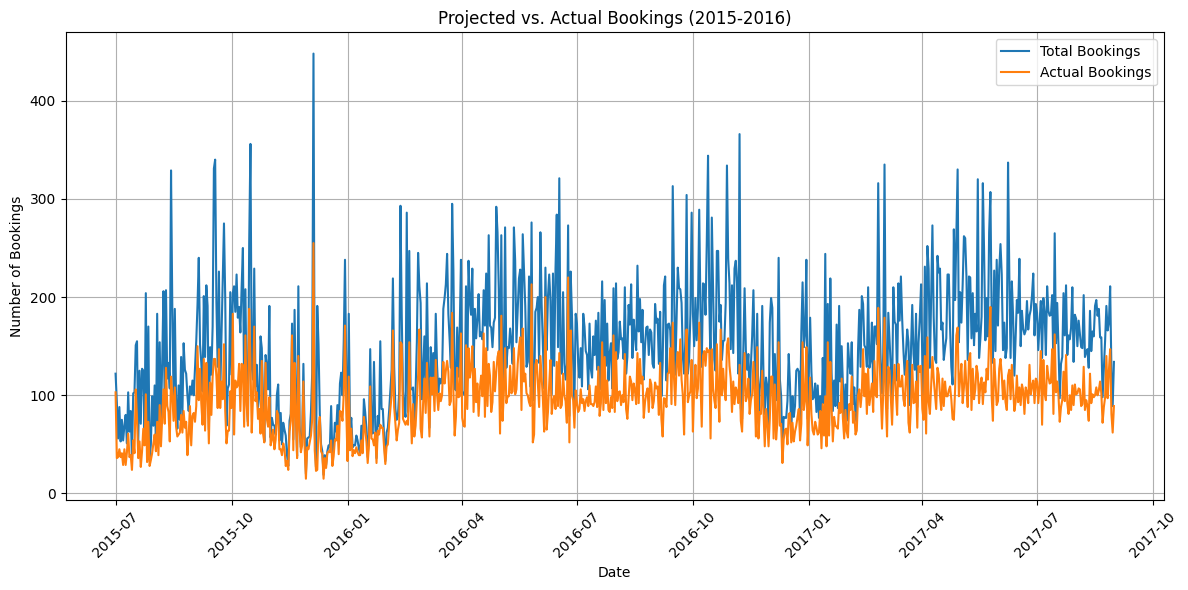
After finding the top five correlators to cancelation, we used a heatmap to show how these variables might interact with each other.



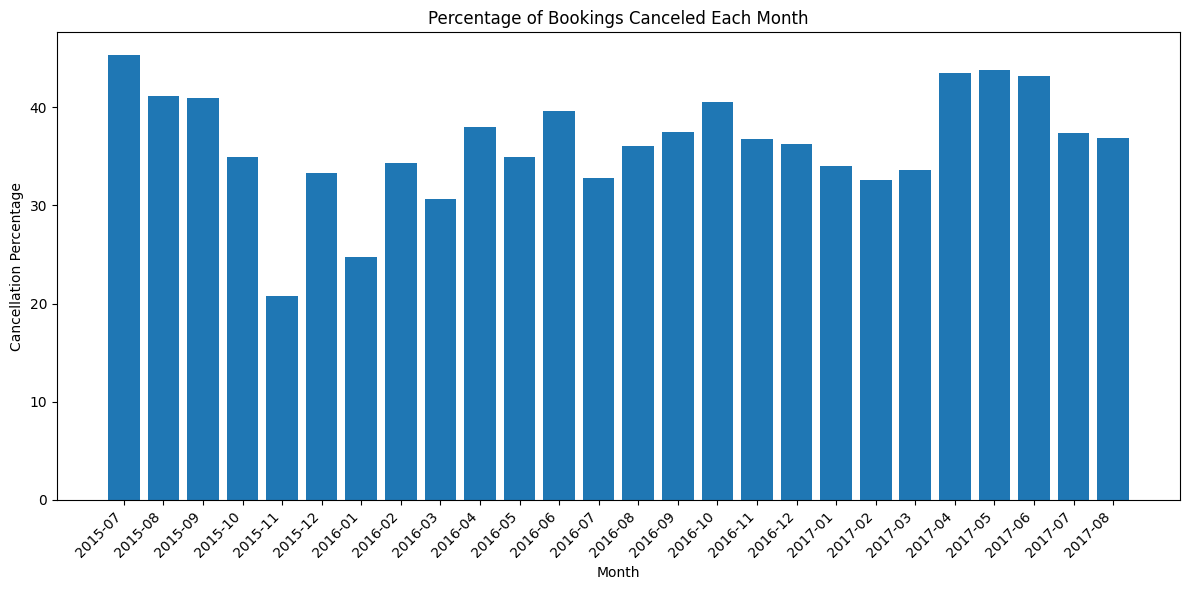
The answer was they interacted relatively little, showing high independence with the highest correlation being .38 between deposit and lead time.

4. Additional Preliminary Analysis

We also calculated the rolling bookings for each date based on cancellation status. We plotted the total number of bookings for each date, as well as the canceled bookings.



While this information is useful, this particular format doesn’t tell us much about how many bookings were canceled beyond “a fair amount”. To improve readability, we next graphed what percentage of each month’s bookings were canceled, rather than individual dates.



This gives us a baseline idea of what we should expect during the modeling phase. A minimum of 20% of bookings were canceled in a given month, a maximum of 45%, and an average of around 35%.