Hotel Booking Analysis & Prediction

Vikas Ranjan

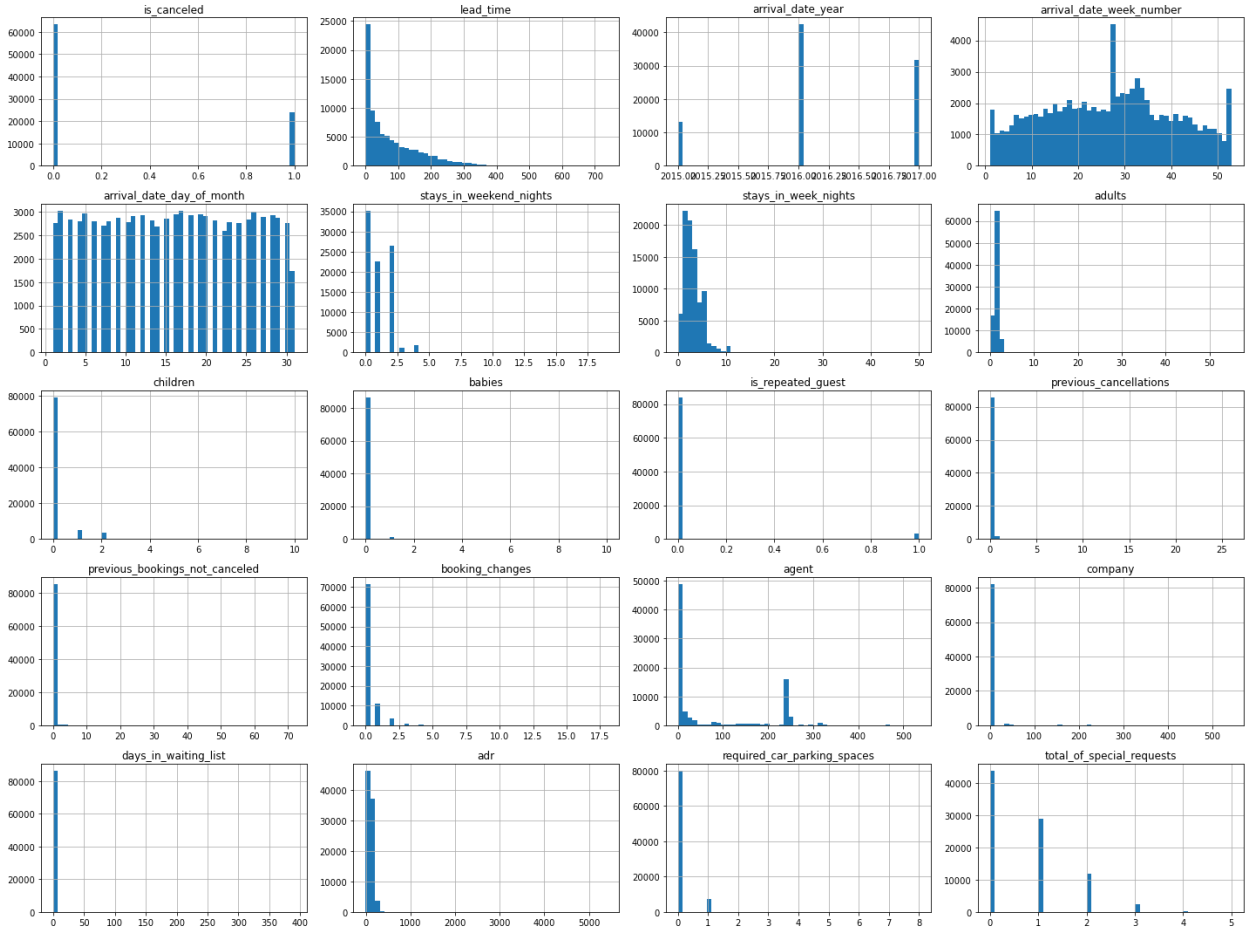
DSC680, Summer 2021

Bellevue University, NE

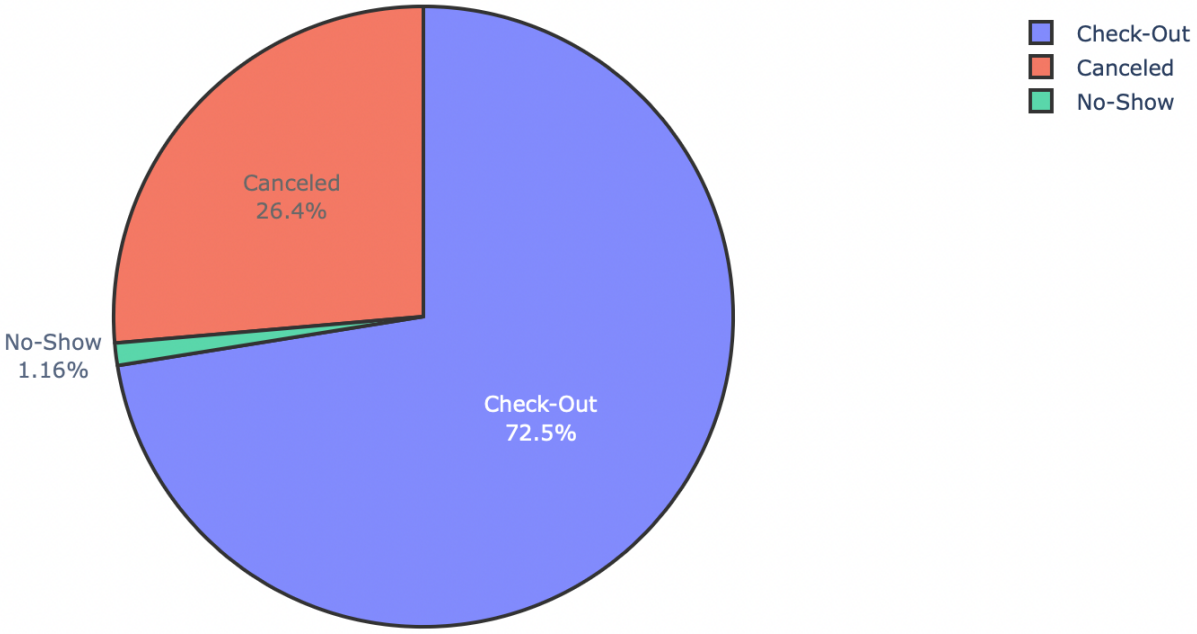
Objective/Questions:

1. Identify distribution of key attributes with regards to booking and cancellations?

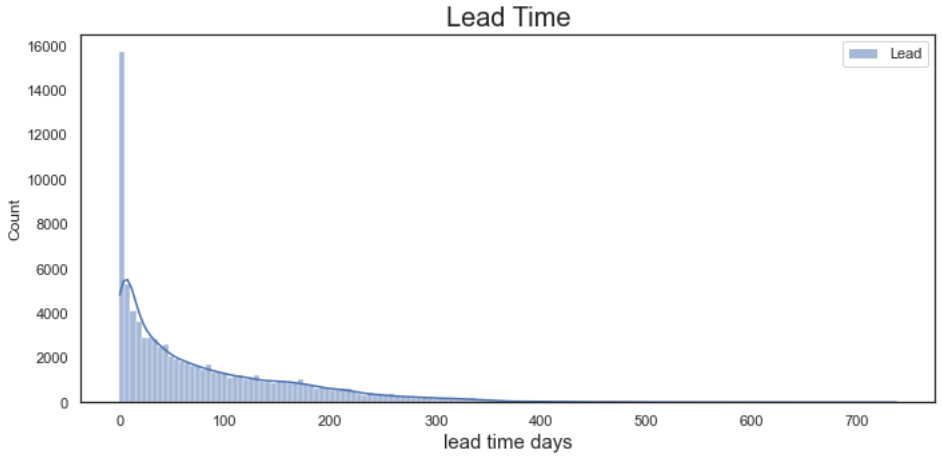
   - Below is the distribution of all the numeric attributes in the dataset.



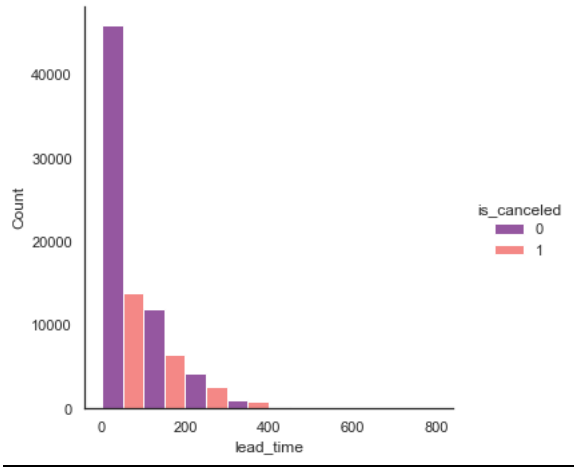
   - Below pie chart shows the distribution of the Reservation Status.



- Below plot shows the distribution and highest concentration points.

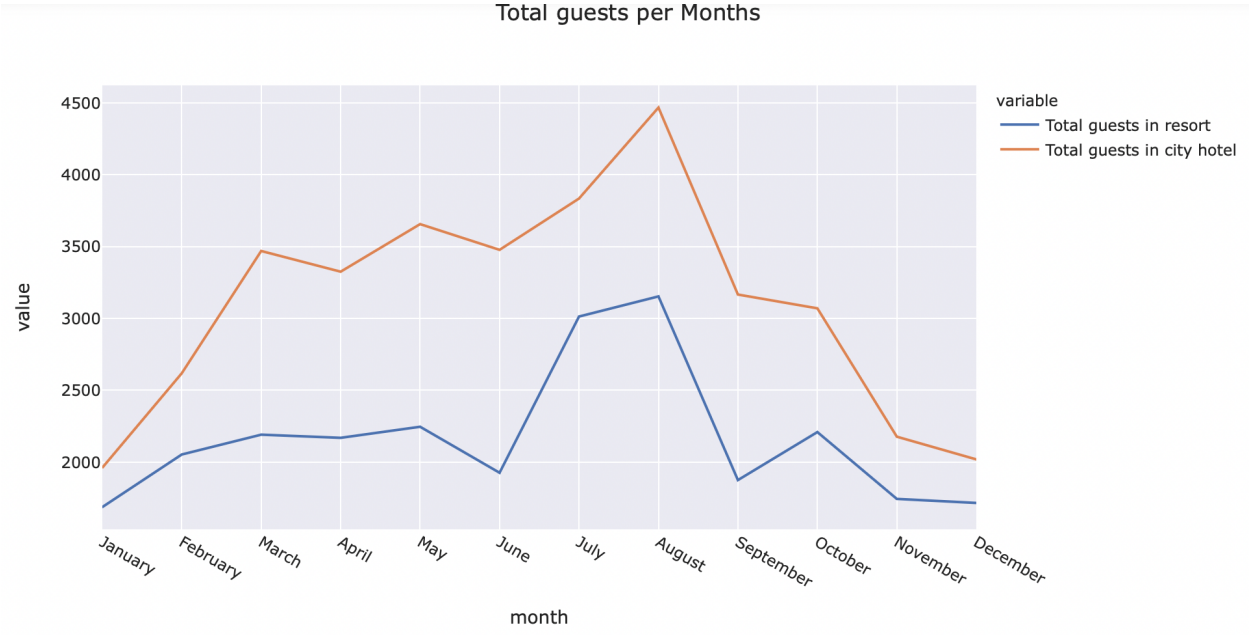


- Below plot shows the distribution of Lead time when compared to cancelled vs non-cancelled bookings.



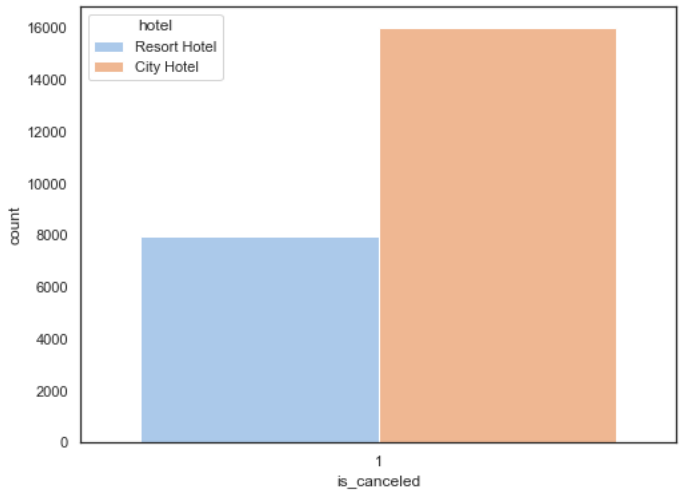
2. How does the City hotel and Resort hotel number of guests per month?

   - Below is the plot showing comparison of total guests per month for city hotels bs resort hotels and here are some observations.
     - The City hotel has more guests during spring and autumn.
     - In July and August there are less visitors.
     - All hotels have the fewest guests during the winter.



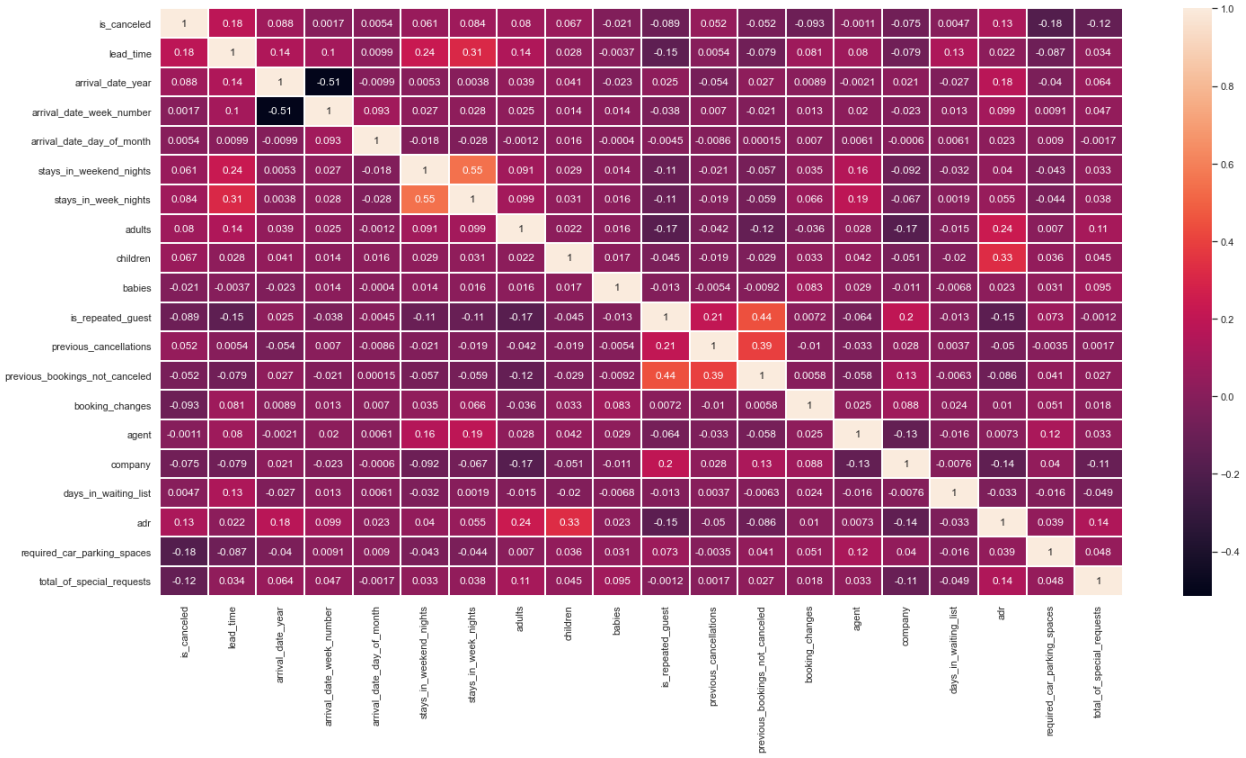
3. Compare the cancellations of resort hotels vs city hotels?

   - Below plot shows the comparison of cancellation of bookings for Resort Hotel and City Hotel.



4. Find out the attributes which are corelated (both positive & negative) with regards to cancellations?

   - Heatmap of the correlation.



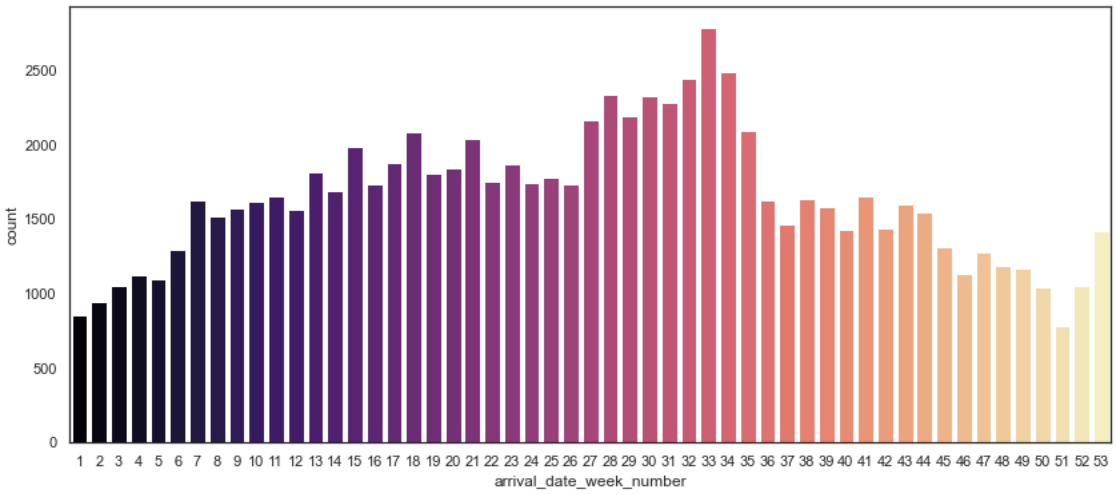

```
is_canceled                       1.000000
lead_time                         0.184515
adr                               0.127233
arrival_date_year                 0.088020
stays_in_week_nights              0.084159
adults                            0.080271
children                          0.067182
stays_in_weekend_nights           0.060992
previous_cancellations            0.051501
arrival_date_day_of_month         0.005449
days_in_waiting_list              0.004710
arrival_date_week_number          0.001691
agent                            -0.001145
babies                           -0.020627
previous_bookings_not_canceled   -0.052178
company                          -0.075314
is_repeated_guest                -0.088764
booking_changes                  -0.093236
total_of_special_requests        -0.120794
required_car_parking_spaces      -0.184456
Name: is_canceled, dtype: float64
```

Correlations Analyis - is_canceled:

- The strongest positive correlations are with lead_time and adr.
- The strongest negative correlations are with total_of_special_requests, required_car_parking_spaces and booking_changes.

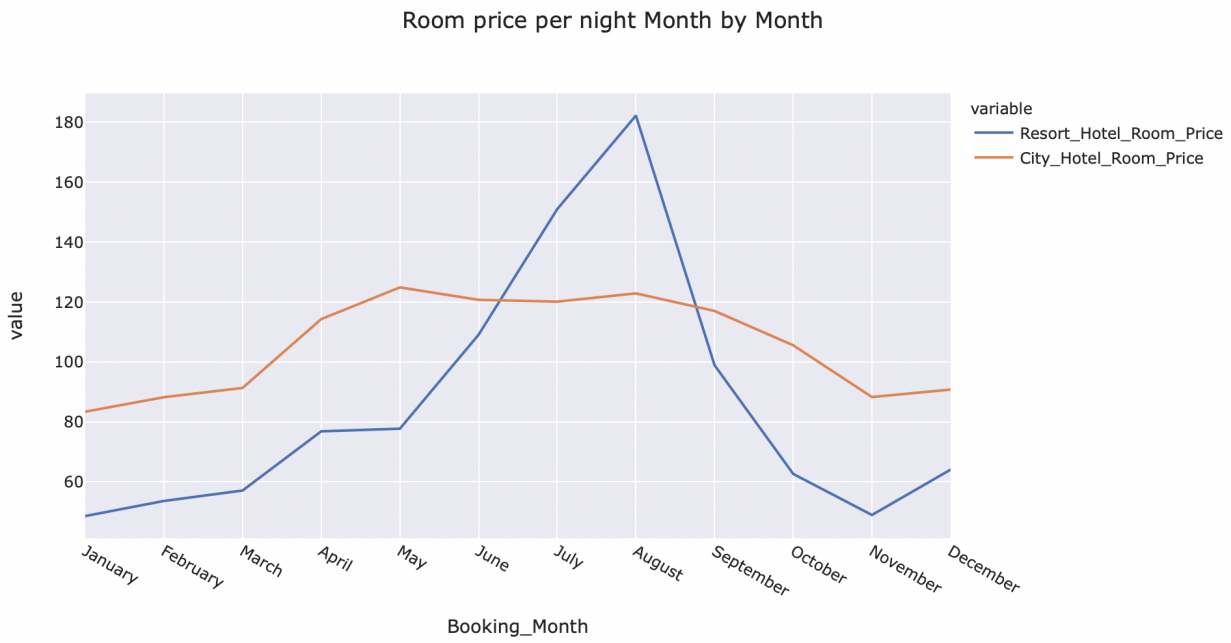
5. Find out insights from distribution of booking over the weeks of the year?

   - Below plot shows the distribution of bookings over the weeks of the year. As we can see, highest number of bookings and guests are between weeks 27 and week 35.



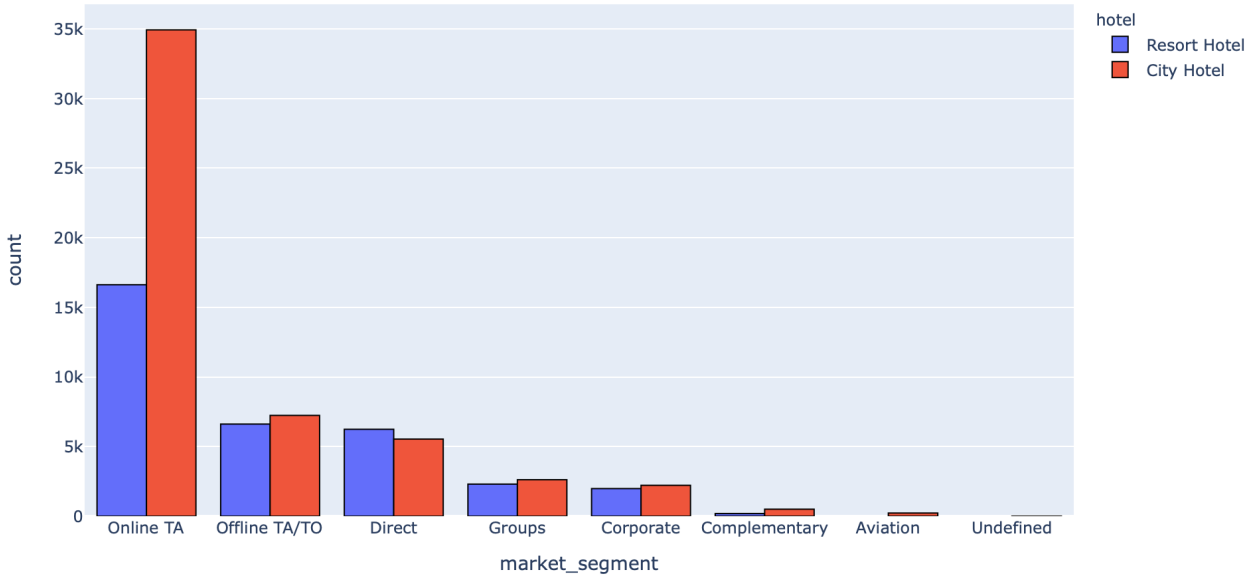
6. Compare and contrast the comparison of Room price per night Month by Month between city hotel and Resort hotel?

   - Below plot shows the comparison of booking price of room over a period of time and as we can see that average price of the resort hotel rooms is more than city hotel rooms. And resort hotel rooms are much expensive during summer months.



7. Analyze the distribution of Market Segment in different Hotel Types?

   - Below is the distribution and here we can see that most of the bookings for Online TA is for City hotel rooms.



8. Identify and perform encoding of categorical.

   - Below are the unique values in the categorical attributes.

```
hotel:
['Resort Hotel' 'City Hotel']

meal:
['BB' 'FB' 'HB' 'SC' 'Undefined']

market_segment:
['Direct' 'Corporate' 'Online TA' 'Offline TA/TO' 'Complementary' 'Groups'
 'Undefined' 'Aviation']

distribution_channel:
['Direct' 'Corporate' 'TA/TO' 'Undefined' 'GDS']

reserved_room_type:
['C' 'A' 'D' 'E' 'G' 'F' 'H' 'L' 'B']

deposit_type:
['No Deposit' 'Refundable' 'Non Refund']

customer_type:
['Transient' 'Contract' 'Transient-Party' 'Group']

year:
[2015 2014 2016 2017]

month:
[ 7  5  4  6  3  8  9  1 11 10 12  2]

day:
[ 1  2  3  6 22 23  5  7  8 11 15 16 29 19 18  9 13  4 12 26 17 10 20 14
 30 28 25 21 27 24 31]
```

- Below is the snapshot of categorical attributes after encoding.

	hotel	meal	market_segment	distribution_channel	reserved_room_type	deposit_type	customer_type	year	month	day
0	0	0	0	0	0	0	0	0	7	1
1	0	0	0	0	0	0	0	0	7	1
2	0	0	0	0	1	0	0	0	7	2
3	0	0	1	1	1	0	0	0	7	2
4	0	0	2	2	1	0	0	0	7	3

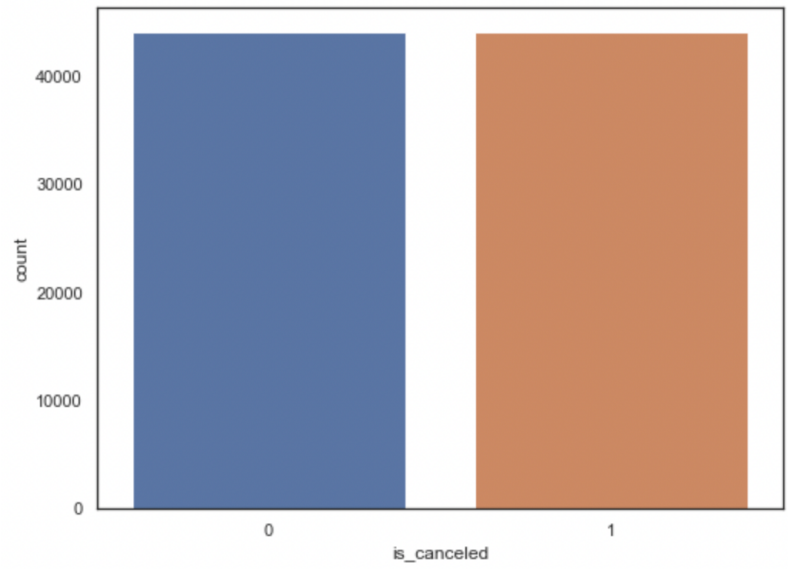
9. Validate if the dataset is imbalanced for applying a classification model? If yes, select a method to overcome this challenge.

   - Comparison of records with cancelled bookings vs non cancelled bookings.

```
bookings_df['is_canceled'].value_counts()
```

```
0    63221
1    24009
Name: is_canceled, dtype: int64
```

   As we can see above, this is a moderately imbalanced dataset. Therefore, we would need to handle that using random oversampling.



10. Identify which classification model will be best suited to predict a booking cancellation?

   - Below is the comparison of the implemented models in terms of accuracy. Here we can see that Random Forest classification model has the best accuracy score followed by Decision Tree classification model when applied on this dataset.

	Model	Score
1	Random Forest Classifier	0.937789
3	Decision Tree Classifier	0.929115
2	KNN	0.816730
0	Logistic Regression	0.680041