FINAL PROJECT BLOG

CHRISTIAN PARK

INTRODUCTION

The dataset that I chose pertains to data from a city hotel and a resort hotel and includes information such as:

- Time of booking
- Length of stay
- # of adults, children
- Whether or not reservation was cancelled
- Etc.

- For this project, I hope to create a predictive model that estimates cancellation rates for these hotels and my target variable is ,'is_cancelled'.

DATA WRANGLING & CLEANING

When initially observing the dataset, I checked for missing values and noticed that 'Children', 'Country', 'agent', and 'company' were the only columns with missing values and therefore had to replace the null values.

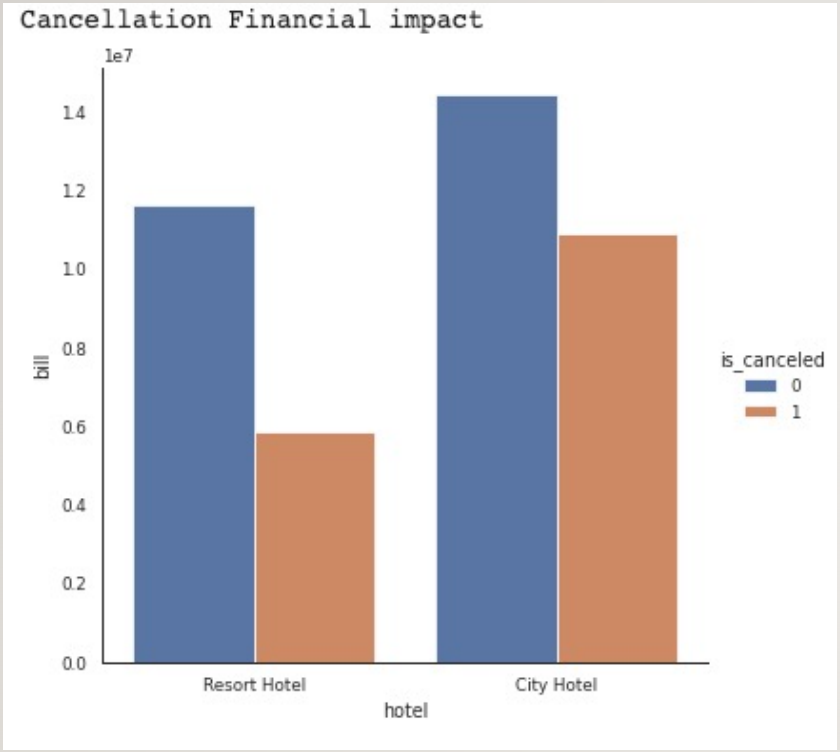
```
booking_data.isnull().sum()
```

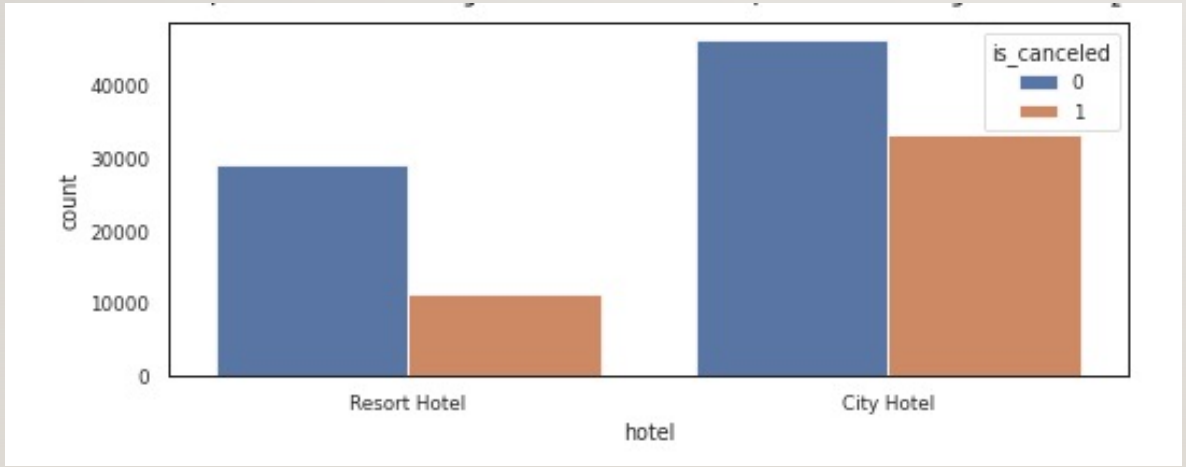
```
hotel                              0
is_canceled                        0
lead_time                          0
arrival_date_year                  0
arrival_date_month                 0
arrival_date_week_number           0
arrival_date_day_of_month          0
stays_in_weekend_nights            0
stays_in_week_nights               0
adults                             0
children                           4
babies                             0
meal                               0
country                          488
market_segment                     0
distribution_channel               0
is_repeated_guest                  0
previous_cancellations             0
previous_bookings_not_canceled     0
reserved_room_type                 0
assigned_room_type                 0
booking_changes                    0
deposit_type                       0
agent                          16340
company                       112593
days_in_waiting_list               0
customer_type                      0
adr                                0
required_car_parking_spaces        0
total_of_special_requests          0
reservation_status                 0
reservation_status_date            0
dtype: int64
```

EDA

WHAT HOTELS SUFFER THE MOST FROM CANCELLATIONS?

- Each row in the dataset refers to a booking, however, one row may be related to one person while another may be for several which can also apply to nights stayed. Therefore, I created some different columns that record total stay nights and the number of booked people per row
- I found that overall, 44224 bookings were canceled, accounting for 37 percent of booked stays

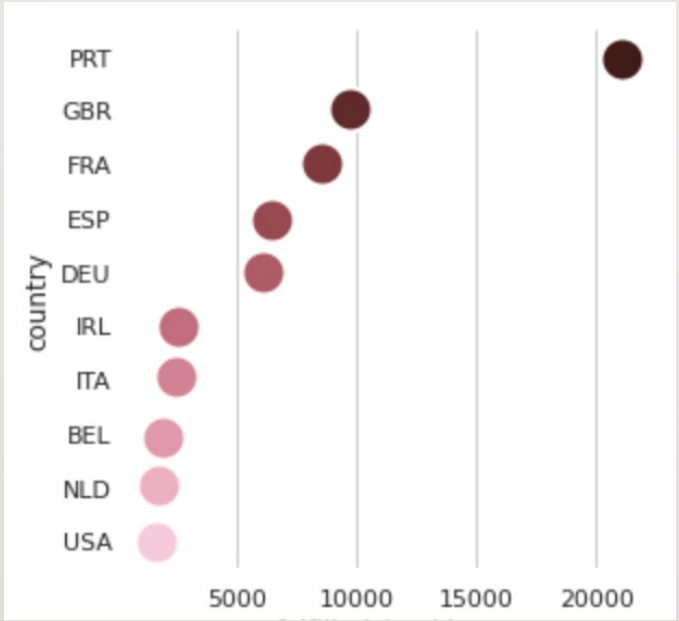


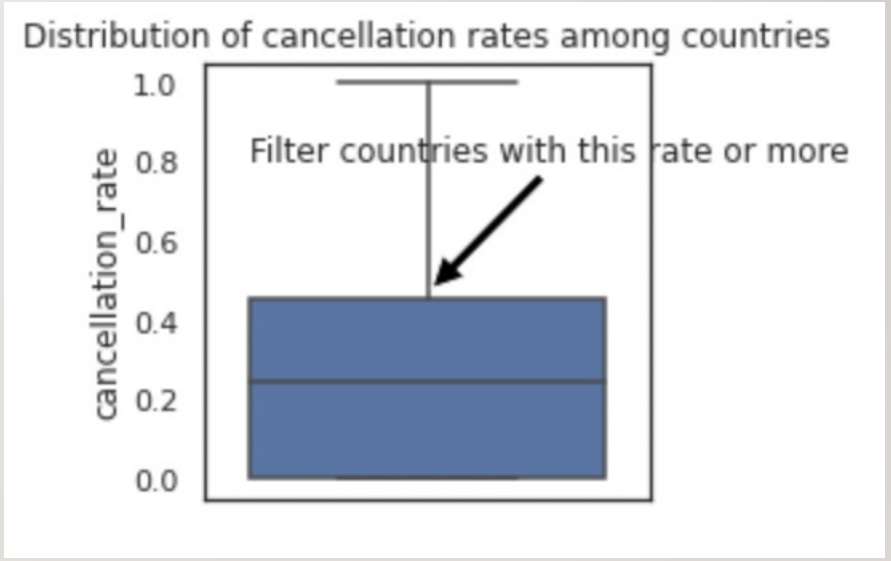


These plots depict how the majority of canceled bookings are in 'city hotels' even though both establishments suffer from cancellations to a relatively similar degree. Thus, we can concur that 'City Hotels' suffer a slight bit more compared to 'Resort Hotels'.

DO PEOPLE FROM A CERTAIN COUNTRY CANCEL MORE THAN OTHERS?

- I wanted to understand the distribution and relationship between a person's origin and how frequently they cancel their reservations. These plots depict the distribution and per country rate of cancellations. The distribution determined that the 3rd quartile is a cancellation rate of 45%





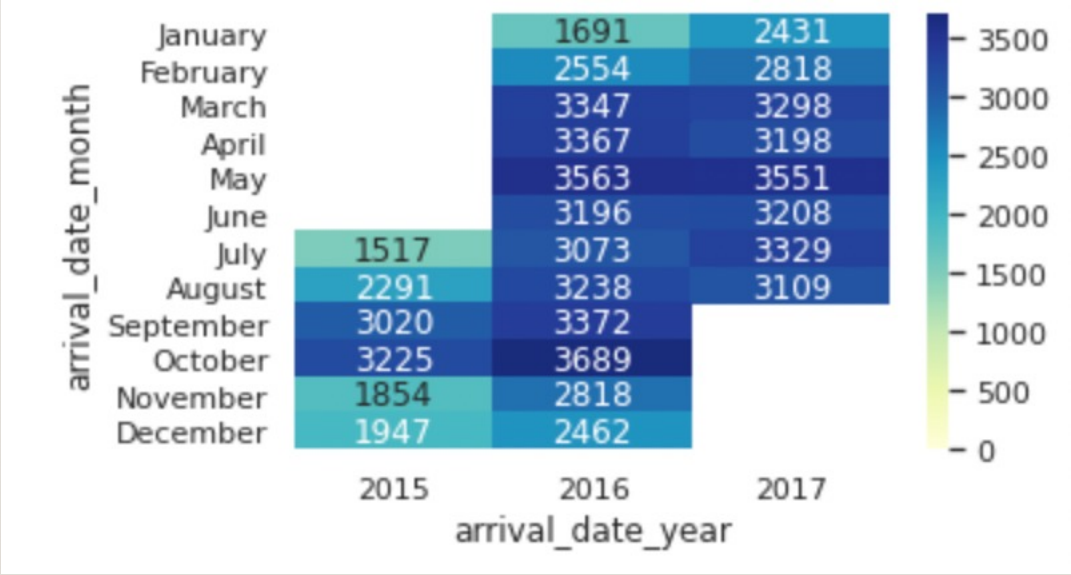
COUNTRIES CONT.

Ultimately, I found that the top 3 countries in terms of cancelling reservations are : Portugal, China and Angola

	is_canceled	stay_nights	pax	booking_count	cancellation_rate
country					
PRT	27519	141654	89599.0	48590	0.566351
CHN	462	2642	2021.0	999	0.462462
AGO	205	2938	647.0	362	0.566298

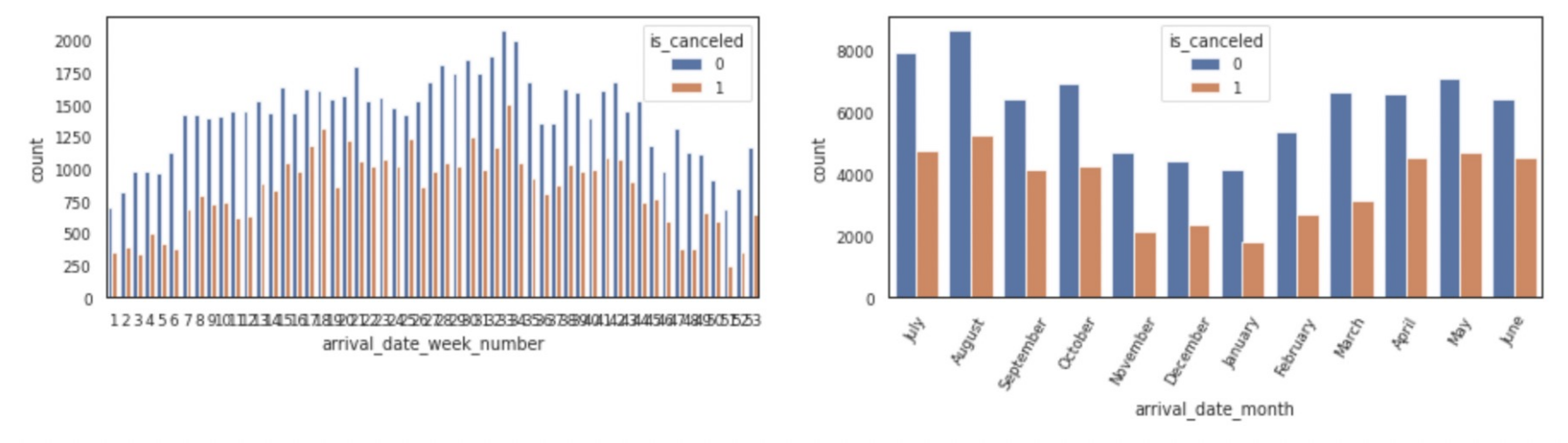
WHAT TIME PERIOD HAS THE PEAK NUMBER OF BOOKINGS? IS THERE SEASONALITY?

- I addressed the seasonality of bookings by finding the months during which the bookings peak.

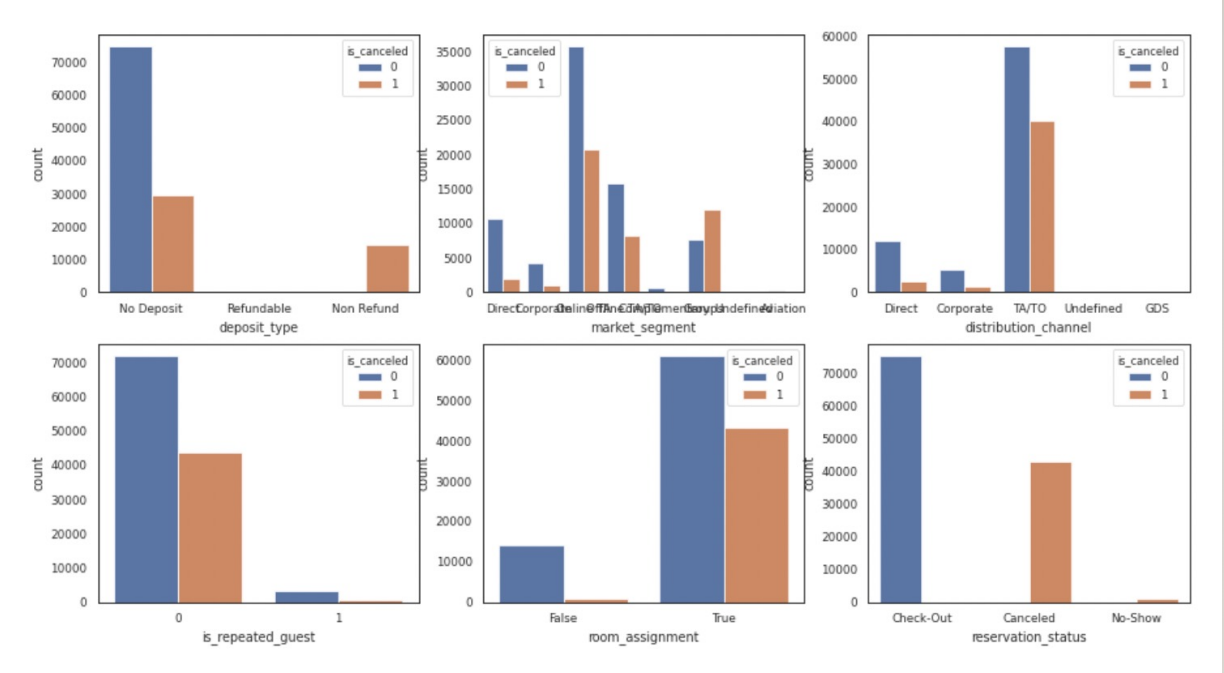


WHAT FACTORS ARE MOST CORRELATED WITH BOOKING CANCELLATION?

- We can infer that there is little correlation between any numerical feature and the target label 'is_cancelled'
- By previewing each feature by itself while counting the fulfilled bookings vs the cancelled ones, we can detect some patterns. In fact, most of the features have a homogenous distribution of cancellations among the unique values, which is a clear indication that there is little chance for the feature to have a correlation with the label to predict. The columns '*arrival_date_month*' and '*arrival_date_week_number*' illustrate this effect



- In fact, we deduced that these are the columns with the highest correlation with the label to predict:
  - In regards to 'deposit_type': 'Non Refund' bookings are canceled almost all the timeAlso, the cancellation rate varies a lot for different values of 'market_segment' and 'distribution_channel'. Repeated guests cancel their bookings a lot fewer than non-repeated guests do. As for the last attribute, there is a perfect connection between the values. For each row that has 'reservation_status' set to 'Check-Out', the label 'is_canceled' is null, and is equal to '1' on the other cases. So this feature will be left out during the training.



ML MODELS

MODEL PERFORMANCES

- DECISION TREE CLASSIFIER
  - ACCURACY: .813
- LOGISTIC REGRESSION
  - ACCURACY: .776
- BAGGING CLASSIFIER
  - ACCURACY: .777
- RANDOM FOREST
  - ACCURACY: .815
- ADA BOOST
  - ACCURACY: .886

CONCLUSION

The best model to predict the outcomes of bookings was **ADA Boosting** on Decision trees, which yields a score as high as 85%.

I found that the model of 88.7% accuracy only helps in predicting revenue to a +/- of 25% and that further statistical testing would be required to narrow down that number. Furthermore, I found that another use of the prediction model would be to intentionally overbook the establishments in peak season and found that the optimal percentage would be 147%.