Hotel Booking Analysis - report

April 18, 2023

1 Hypothesis

1. More cancellations occur when prices are higher.
2. When there is a longer waiting list, customers tend to cancel more frequently.
3. The majority of clients are coming from offline travel agents to make their reservations.

2 Data Analysis and Visualizations

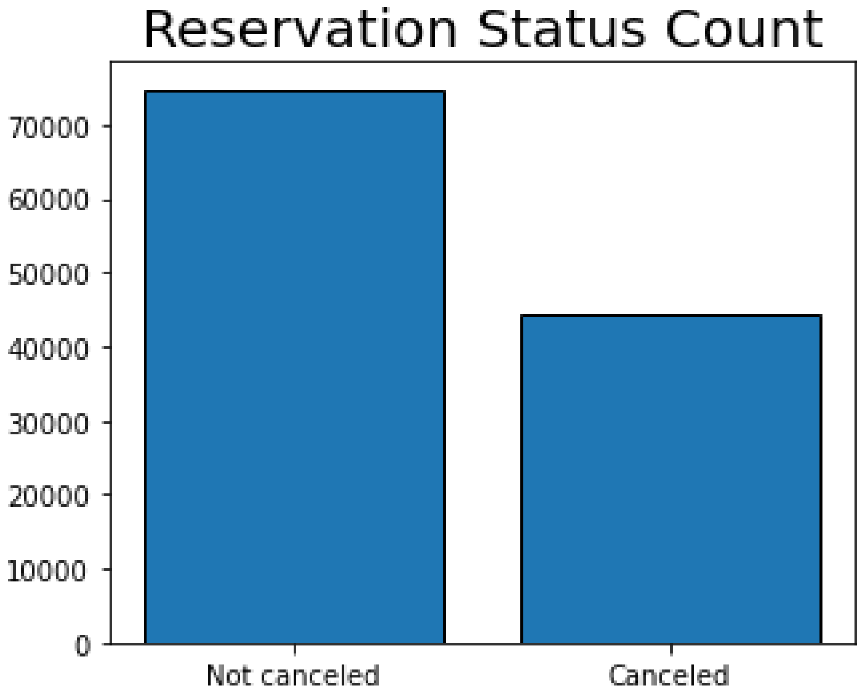
[28]:

```
canceled_perc = df['is_canceled'].value_counts(normalize = True)
print(canceled_perc)
```

```
0    0.628653
1    0.371347
Name: is_canceled, dtype: float64
```

[39]:

```
plt.figure(figsize = (5,4))
plt.title('Reservation Status Count', fontsize = 20)
plt.bar(['Not canceled', 'Canceled'], df['is_canceled'].value_counts(),␣
 ↪edgecolor = 'k', width = 0.8)
plt.show()
```



The accompanying bar graph shows the percentage of reservations that are cancelled and those that are not. It is obvious that there are still a significant number of reservations that have not been cancelled. There are still 37% of clients who cancelled their reservation, which has an important impact on the hotels' earnings.

```
[33]: resort_hotel = df[df['hotel'] == 'Resort Hotel']
      resort_hotel['is_canceled'].value_counts(normalize = True)
```

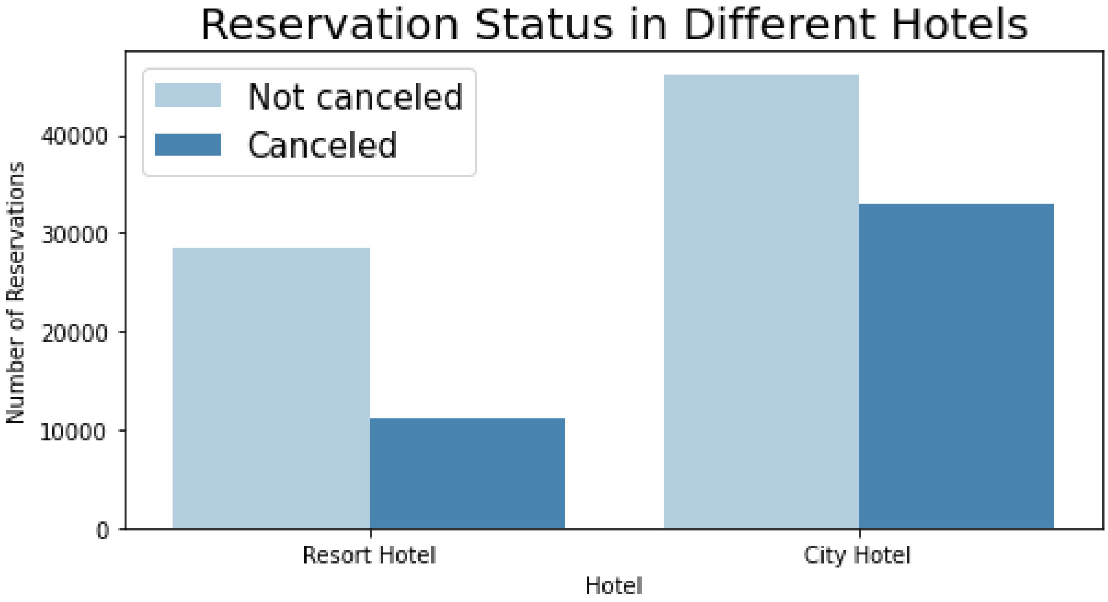
```
[33]: 0    0.72025
      1    0.27975
      Name: is_canceled, dtype: float64
```

```
[34]: city_hotel = df[df['hotel'] == 'City Hotel']
      city_hotel['is_canceled'].value_counts(normalize = True)
```

```
[34]: 0    0.582918
      1    0.417082
      Name: is_canceled, dtype: float64
```

```
[47]: plt.figure(figsize = (8,4))
      ax1 = sns.countplot(x = 'hotel', hue = 'is_canceled', data = df, palette =␣
       ↪'Blues')
      legend_labels,_ = ax1.get_legend_handles_labels()
      ax1.legend(bbox_to_anchor = (1,1))
```

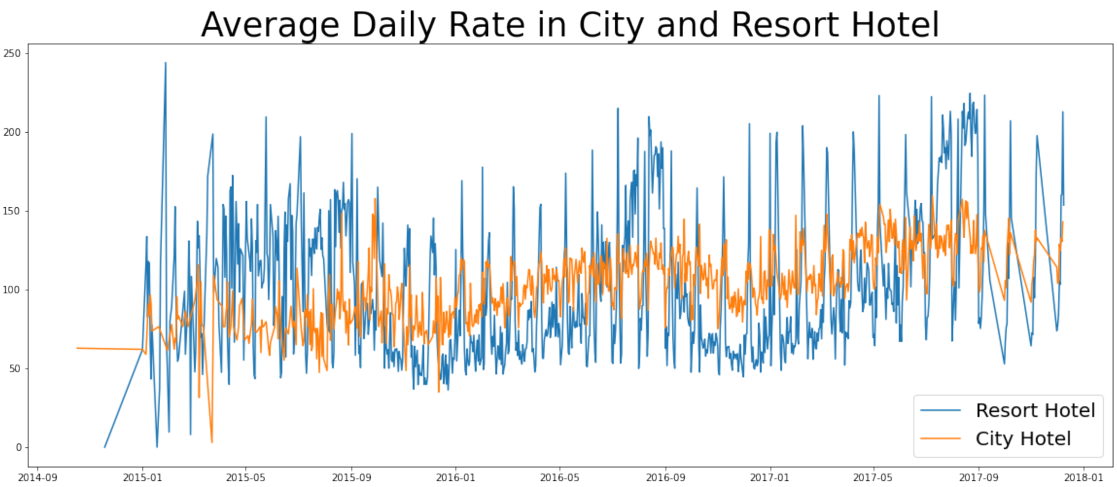
```
plt.title('Reservation Status in Different Hotels', size = 20)
plt.xlabel('Hotel')
plt.ylabel('Number of Reservations')
plt.legend(['Not canceled', 'Canceled'], fontsize = 15)
plt.show()
```



In comparison to resort hotels, city hotels have more bookings. It is possible that resort hotels are more expensive than those in cities.

Interestingly, the disparity between the percentage of non-cancellations and cancellations is far greater for resort hotels. Only 27% of bookings at these hotels were cancelled, compared to 41% at city hotels. This may also have something to do with price, as it is much easier to cancel a cheaper booking in a city hotel and expect a lower financial loss than in the case of resort hotels, where the reservation and deposit alone are much higher.

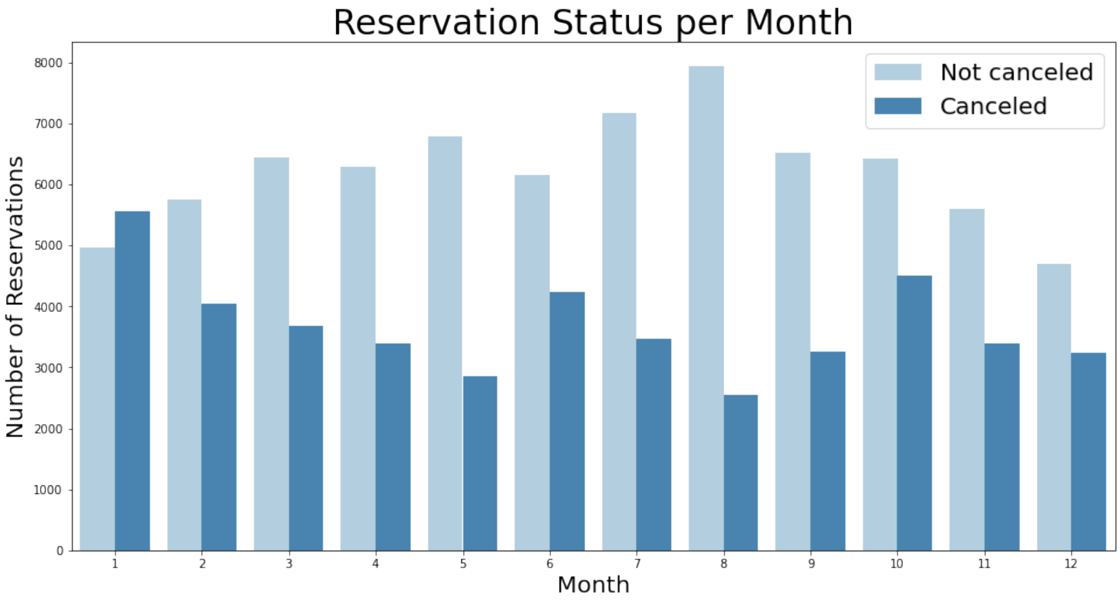
[55]:
```
plt.figure(figsize = (20,8))
plt.title('Average Daily Rate in City and Resort Hotel', fontsize = 35)
plt.plot(resort_hotel.index, resort_hotel['adr'], label = 'Resort Hotel')
plt.plot(city_hotel.index, city_hotel['adr'], label = 'City Hotel')
plt.legend(fontsize = 20)
plt.show()
```



The line bar above shows that, on certain days, the average daily rate for a city hotel is less than that of a resort hotel, and on other days, it is even less. It goes without saying the weekends and holidays may see a rise in resort hotel rates.

[53]:

```
df['month'] = df['reservation_status_date'].dt.month
plt.figure(figsize = (16,8))
ax1 = sns.countplot(x = 'month', hue = 'is_canceled', data = df, palette =␣
 ↪'Blues')
plt.title('Reservation Status per Month', size = 30)
plt.xlabel('Month', fontsize = 20)
plt.ylabel('Number of Reservations', fontsize = 20)
plt.legend(['Not canceled', 'Canceled'], fontsize = 20)
plt.show()
```

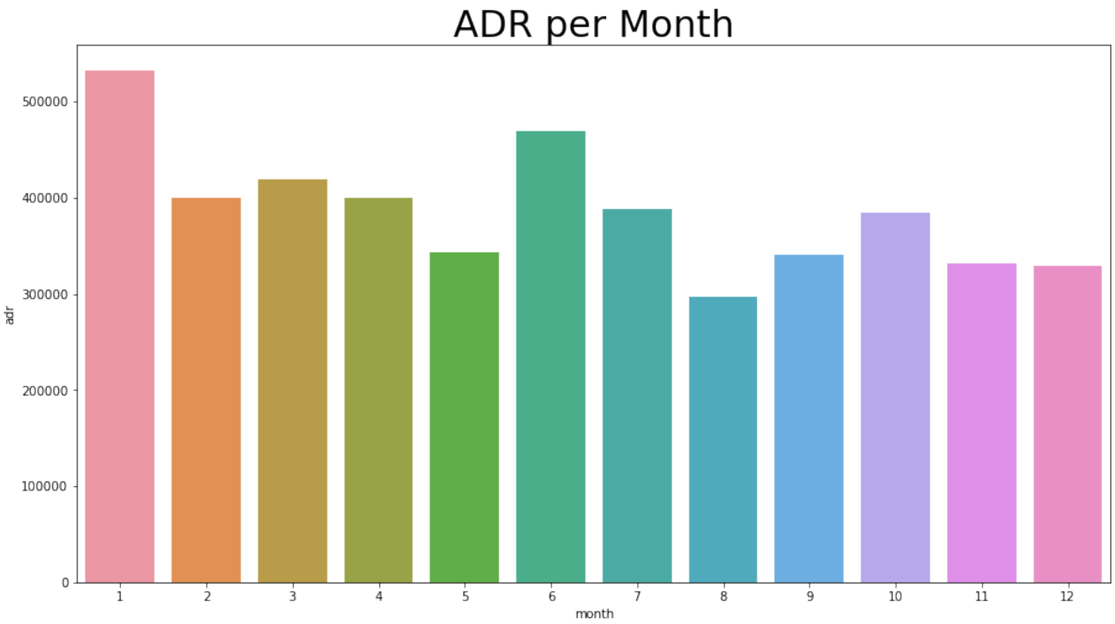


As we can see on the bar graph above, both the month with the highest number of non-canceled reservations is August, what's more that the month with the highest number of all reservations. Importantly, this is the month with the highest ratio between non-cancellations and cancellations. This is not much of a surprise as it is the holiday season and this is when most people decide to go on holiday.

The month with the highest number of cancellations is January. Notably, it is the only month in which the number of cancelled bookings is higher than the number of non-cancelled bookings.

[61]:

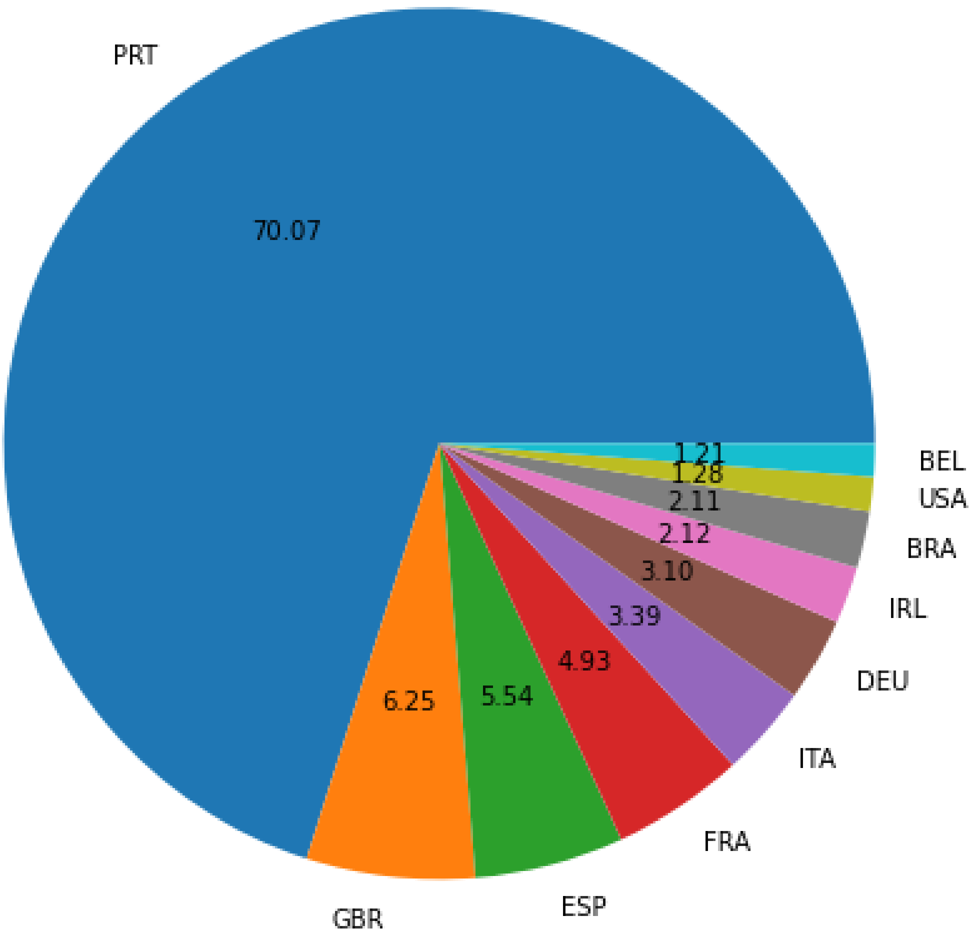
```python
plt.figure(figsize = (15,8))
plt.title('ADR per Month', fontsize = 30)
sns.barplot('month', 'adr', data = df[df['is_canceled'] == 1].groupby('month')[['adr']].sum().reset_index())
plt.show()
```



The bar graph demonstrates that cancellations are most common when prices are higher.

[62]:

```python
canceled_data = df[df['is_canceled'] == 1]
top_10_country = canceled_data['country'].value_counts()[:10]
plt.figure(figsize = (8,8))
plt.pie(top_10_country, autopct = '%.2f', labels = top_10_country.index)
plt.show()
```



Portugal is the clear leader in terms of the percentage of cancelled bookings with as many as 70%.

```
[65]: canceled_data['market_segment'].value_counts(normalize = True)
```

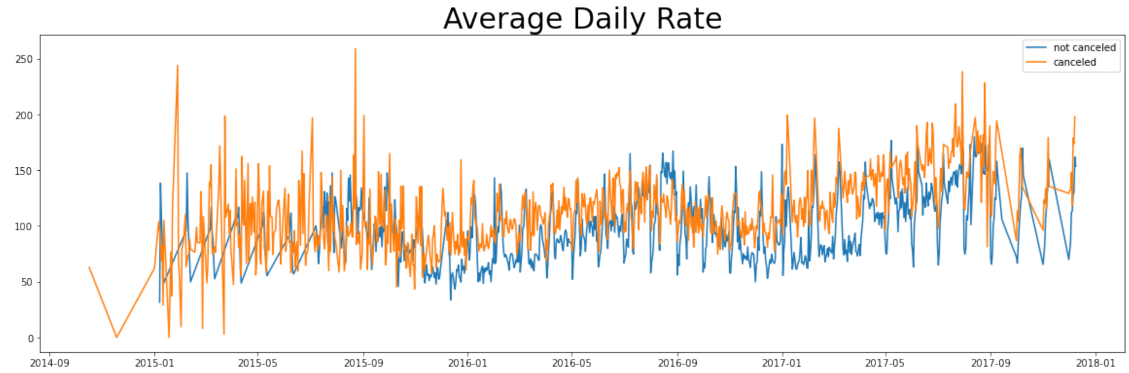
```
[65]: Online TA        0.469696
      Groups           0.273985
      Offline TA/TO    0.187466
      Direct           0.043486
      Corporate        0.022151
      Complementary    0.002038
      Aviation         0.001178
      Name: market_segment, dtype: float64
```

Around 47% if the clients come from online agencies, whereas 27% come from groups. Only 4% of clients books hotels directly by visiting them and making reservations. This comes as no surprise, as hotels are usually booked not only in a foreign city or country, but thousands of kilometres away from the place of residence, so it is logical that in such situations booking on the spot is pointless. These 4% probably account for the vast majority of hotel bookings in the place of residence or not far away.

```
[69]: canceled_df_adr = canceled_data.groupby('reservation_status_date')[['adr']].
       ↪mean()
      canceled_df_adr.reset_index(inplace = True)
      canceled_df_adr.sort_values('reservation_status_date', inplace = True)

      not_canceled_data = df[df['is_canceled'] == 0]
      not_canceled_df_adr = not_canceled_data.
       ↪groupby('reservation_status_date')[['adr']].mean()
      not_canceled_df_adr.reset_index(inplace = True)
      not_canceled_df_adr.sort_values('reservation_status_date', inplace = True)

      plt.figure(figsize =(20,6))
      plt.title('Average Daily Rate', fontsize = 30)
      plt.plot(not_canceled_df_adr['reservation_status_date'],␣
       ↪not_canceled_df_adr['adr'], label = 'not canceled')
      plt.plot(canceled_df_adr['reservation_status_date'], canceled_df_adr['adr'],␣
       ↪label = 'canceled')
      plt.legend()
      plt.show()
```



As seen in the graph, reservations are cancelled when the average daily rate is higher than when it is not cancelled. It clearly proves all the above anlysis, that the higher price leads to higher cancellation.

3 Suggestions

1. Cancellation rates rise as the price does. In order to prevent cancellations of reservations, hotels could work on their pricing strategies and try to lower the rates for specific hotels based

on locations. They can also provide some discounts to the customers.

2. The ratio of cancellation and not cancellation of the resort hotels is higher than the city hotels. So the hotels should provide a reasonable discount on the room prices on weekends or on holidays.
3. In the month of January, hotels can start their campaigns or marketing with a reasonable amount to increase their revenue as the cancellation is the highest in this month.
4. They can also increase the quality of their hotels and services mainly in Portugal, to reduce this high cancellation rate.