Hotel Booking Analysis

Lunjing Yuan

Final project for MSIS 2507 Data Analytics with Python class at Santa Clara University

Data Cleaning

See totally how many NaNs are in the dataset

```
df.isna().sum().sum()
```

```
129425
```

Check NaN again

```
df.isna().any()
```

```
is_canceled                       False
lead_time                         False
arrival_date_year                 False
arrival_date_month                False
arrival_date_week_number          False
arrival_date_day_of_month         False
stays_in_weekend_nights           False
stays_in_week_nights              False
adults                            False
children                          False
babies                            False
meal                              False
country                           False
market_segment                    False
distribution_channel              False
is_repeated_guest                 False
previous_cancellations            False
previous_bookings_not_canceled    False
reserved_room_type                False
assigned_room_type                False
booking_changes                   False
deposit_type                      False
days_in_waiting_list              False
customer_type                     False
adr                               False
required_car_parking_spaces       False
total_of_special_requests         False
reservation_status                False
reservation_status_date           False
dtype: bool
```

1. Cancellation by market segment ?

Count of each market segment and market cancellation

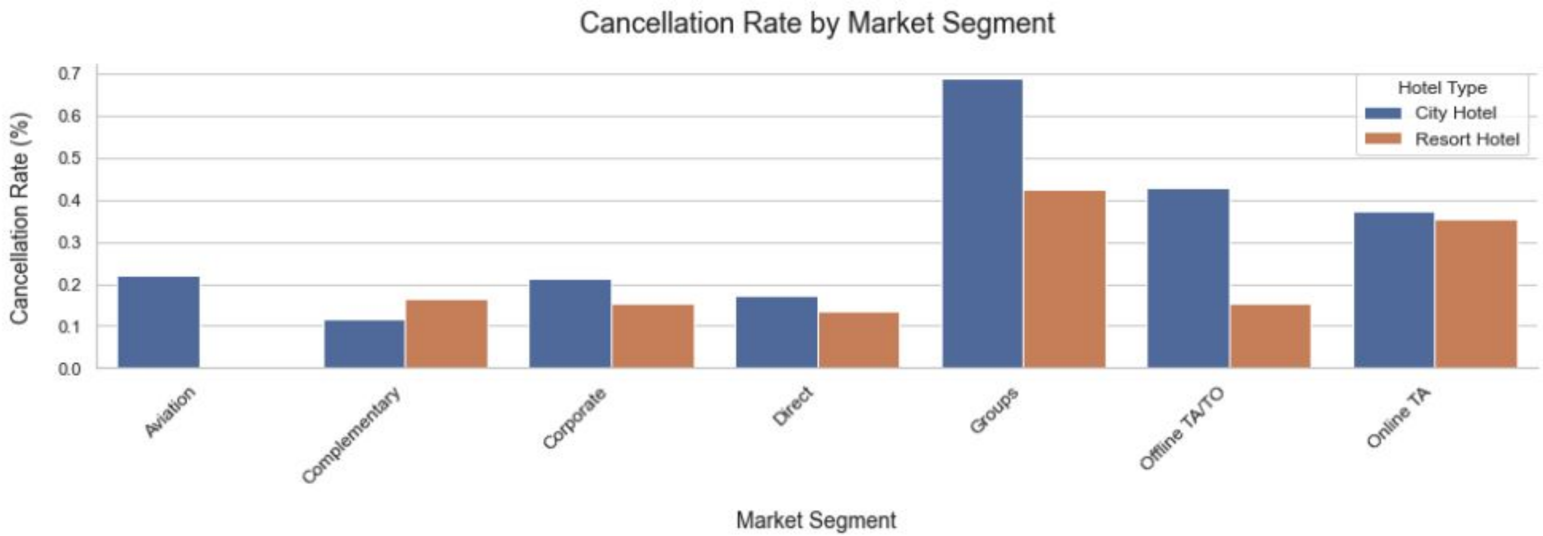
```
total_segment = df.groupby(['hotel','market_segment'])['market_segment'].count()
total_canceled = df.groupby(['hotel','market_segment'])['is_canceled'].sum()
```

Calculation of each market cancellation rate

```
cancel_rate = total_canceled/total_segment
cancel_rate
```

```
hotel         market_segment
City Hotel    Aviation          0.219409
              Complementary     0.118081
              Corporate         0.214668
              Direct            0.173314
              Groups            0.688587
              Offline TA/TO     0.428316
              Online TA         0.373981
              Undefined         1.000000
Resort Hotel  Complementary     0.164179
              Corporate         0.152014
              Direct            0.134807
              Groups            0.423920
              Offline TA/TO     0.152302
              Online TA         0.352417
dtype: float64
```

Finding 1:
The group reservations exhibit the highest cancellation rate



Managerial insights 1: Given the notably elevated cancellation rate among group reservations, we recommend that hotels consider implementing a penalty system, such as rendering hotel prepayments non-refundable.

Which months experience the highest level of activity?

1. Utilize bookings that have been checked-out to obtain the actual count of orders.

2. Employ the month of arrival and hotel types to determine the total number of orders per month.

3. Construct a sub-dataframe that encompasses both City and Resort hotel data to compute the average number of orders per month.

```
df2['resort_hotel'] = (df2.hotel == "Resort Hotel") & (df2.reservation_status == 'Check-Out')
df2['city_hotel'] = (df2.hotel == "City Hotel") & (df2.reservation_status == 'Check-Out')
```

```
total_order.loc[(total_order ["month"] == "July") | (total_order ["month"] == "August"),"orders"] /= 3
total_order.loc[~((total_order ["month"] == "July") | (total_order ["month"] == "August")),"orders"] /= 2
total_order
```

	month	hotel	orders
0	April	City hotel	2007.500000
1	August	City hotel	1793.666667
2	December	City hotel	1196.000000
3	February	City hotel	1532.000000
4	January	City hotel	1127.000000
5	July	City hotel	1594.000000
6	June	City hotel	2183.000000
7	March	City hotel	2036.000000
8	May	City hotel	2289.500000
9	November	City hotel	1348.000000
10	October	City hotel	2168.500000
11	September	City hotel	2145.000000
12	April	Resort hotel	1275.000000
13	August	Resort hotel	1085.666667
14	December	Resort hotel	1008.500000
15	February	Resort hotel	1154.000000
16	January	Resort hotel	934.000000
17	July	Resort hotel	1045.666667
18	June	Resort hotel	1019.000000
19	March	Resort hotel	1286.500000
20	May	Resort hotel	1267.500000
21	November	Resort hotel	988.000000
22	October	Resort hotel	1288.500000
23	September	Resort hotel	1051.000000

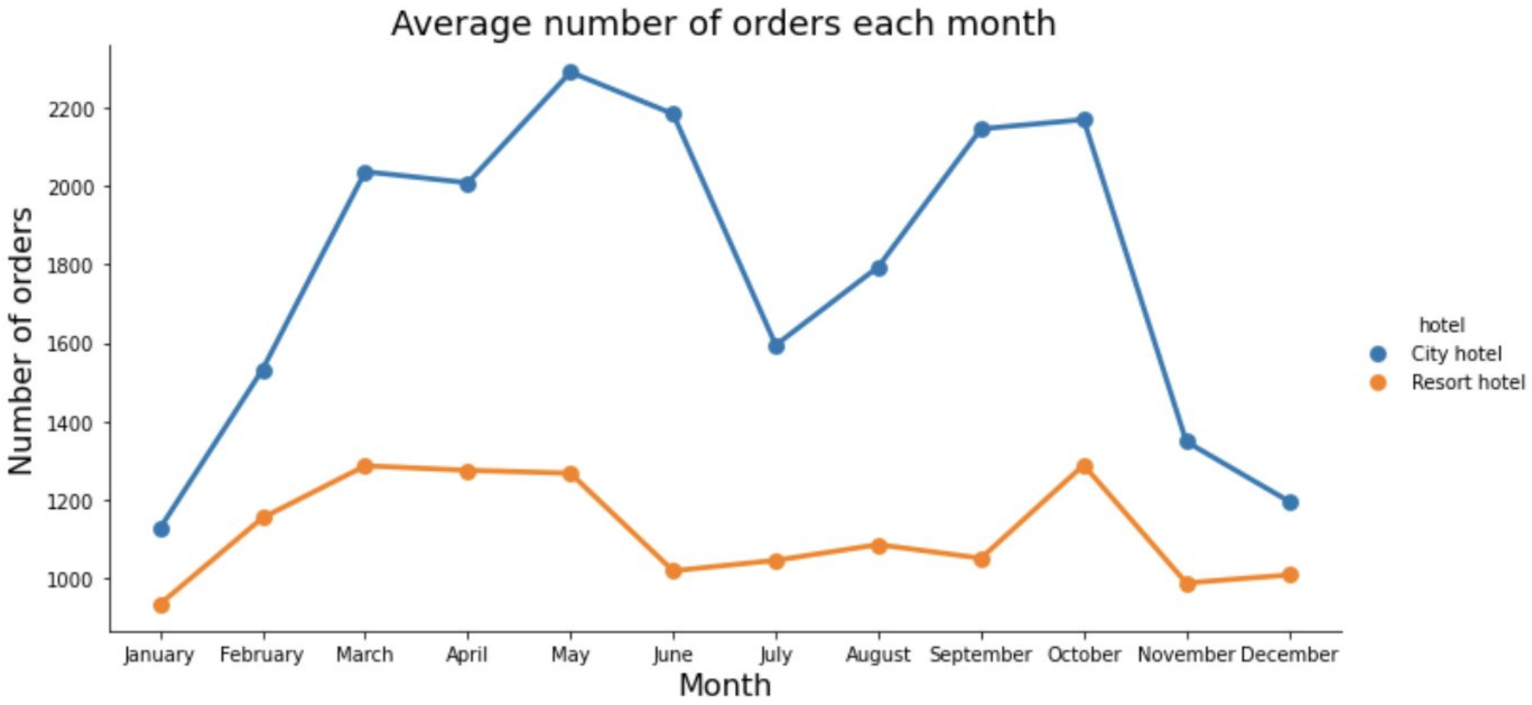
The data is between July 1st 2015 and the August 31st 2017:

2015: 7,8,9,10,11,12

2016: 1,2,3,4,5,6, 7,8,9,10,11,12

2017: 1,2,3,4,5,6, 7,8

Finding 2:
The summer season witnessed a substantial decline in the volume of orders



Managerial insights 2

To offer recommendations, my focus is on identifying the top three countries with the highest booking records.

1. Targeting local leisure travelers can expand the share of bookings within Portugal. Hosting summer events can draw Portuguese guests seeking fun experiences aligned with local culture and seasonal travel motives. To attract domestic Portuguese guests, hotels can host engaging local events like summer barbeque contests and pool parties.

2. Offer discounts or incentives such as free breakfast for orders made in Germany and France during the summer season.

```
country
PRT    48590
GBR    12129
FRA    10415
ESP     8568
DEU     7287
ITA     3766
IRL     3375
BEL     2342
BRA     2224
NLD     2104
Name: country, dtype: int64
```

What factors contribute to cancellations?

i) Get 10 correlated variables and construct a dataframe

ii) Make X and Y then run the Lasso Regression.

iii) Plot graph to validate.

```
canceled_cor = df.corr()['is_canceled']
canceled_cor
```

```
is_canceled                        1.000000
lead_time                          0.293123
arrival_date_year                  0.016660
arrival_date_week_number           0.008148
arrival_date_day_of_month         -0.006130
stays_in_weekend_nights           -0.001791
stays_in_week_nights               0.024765
adults                             0.060017
children                           0.005036
babies                            -0.032491
is_repeated_guest                 -0.084793
previous_cancellations             0.110133
previous_bookings_not_canceled    -0.057358
booking_changes                   -0.144381
days_in_waiting_list               0.054186
adr                                0.047557
required_car_parking_spaces       -0.195498
total_of_special_requests         -0.234658
Name: is_canceled, dtype: float64
```

Finding 3:
Reservation cancellations tend to rise as the lead time becomes longer.

Explore relevant factors

```
regLasso.coef_
```

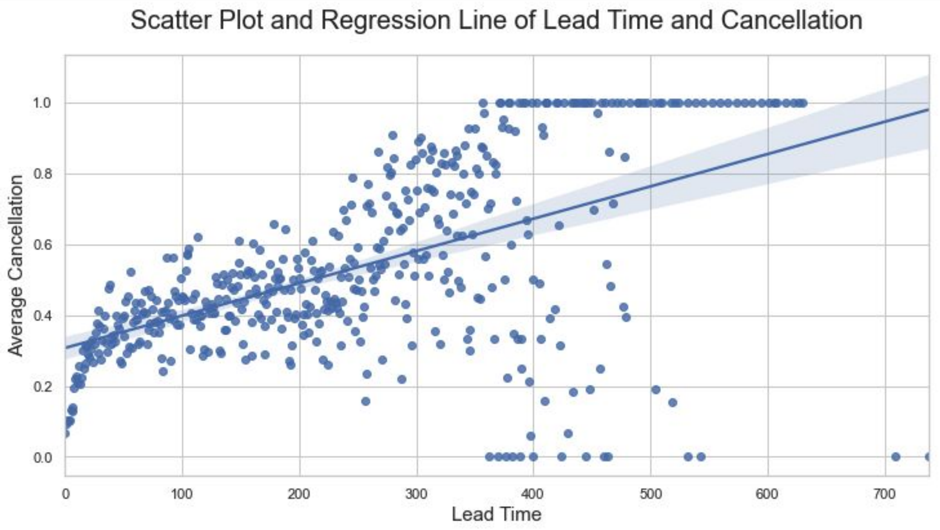
```
array([ 0.00123707, -0.        , -0.        , -0.        ,  0.        ,
       -0.        ,  0.        , -0.        ,  0.        ])
```

```
d = {X.columns[i] : regLasso.coef_[i] for i in range(0,len(X.columns))}
```

```
s = pd.Series(d)
```

```
s[s != 0]
```

```
lead_time    0.001237
dtype: float64
```



Managerial insights 3

To incentivize guests to uphold long lead time reservations, hotels could offer progressive discounts for early booking:

- 10% discount for reservations made 50 days in advance
- 15% discount for reservations made 100 days in advance
- 20% discount for reservations made 150 days in advance