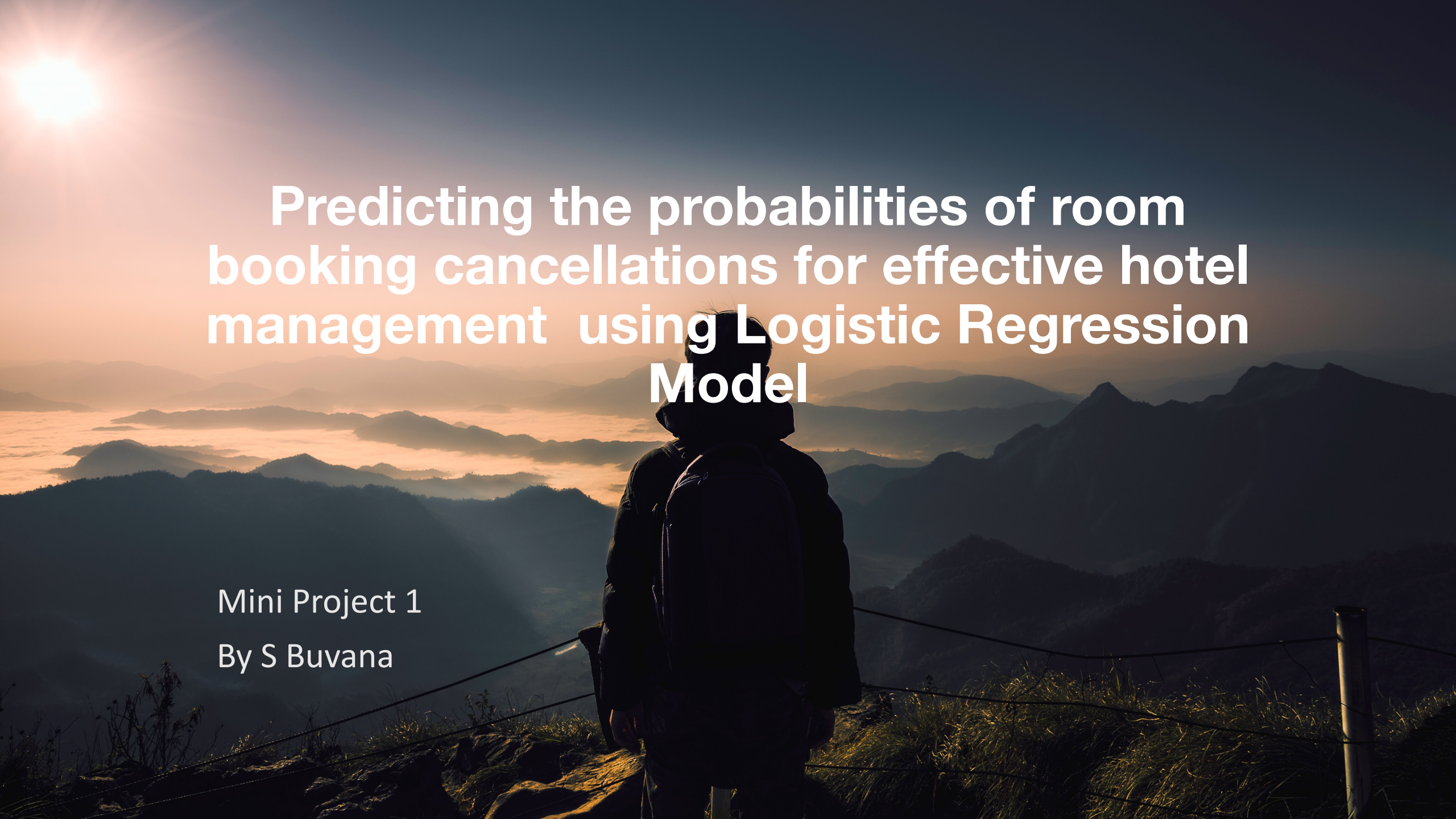


Predicting the probabilities of room booking cancellations for effective hotel management using Logistic Regression Model

Mini Project 1

By S Buvana

Introduction to Dataset: Hotel Bookings Demand

- Dataset contains data about two Hotels; **Resort and City Hotel** based at Portugal
  - Key information such as the types of hotel guests, arrival information, duration of stays, cancellation rates, allocation of room types and lead time
- Dataset entails the information of hotel guests who have booked their stays from **July 2015 to Aug 2017.**
- No identifier columns (customer id/reservation id/personal details)

Load Dataset- Hotel Bookings

In [2]:
```
Hotel_bookings = pd.read_csv("/Users/sbuvana/Desktop/Labs/data/hotel_bookings.csv")
Hotel_bookings.head(10)
```

Out[2]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_i
0	Resort Hotel	0	342	2015	July	27	1	0	
1	Resort Hotel	0	737	2015	July	27	1	0	
2	Resort Hotel	0	7	2015	July	27	1	0	
3	Resort Hotel	0	13	2015	July	27	1	0	
4	Resort Hotel	0	14	2015	July	27	1	0	
5	Resort Hotel	0	14	2015	July	27	1	0	
6	Resort Hotel	0	0	2015	July	27	1	0	
7	Resort Hotel	0	9	2015	July	27	1	0	
8	Resort Hotel	1	85	2015	July	27	1	0	
9	Resort Hotel	1	75	2015	July	27	1	0	

10 rows × 32 columns

Initial Questions

1) How in advance do the hotel guests book their rooms?

2) Who forms the majority of the hotel guest list?

3) Does cancellation occur amongst guests who have paid advances?

4) Which hotel has higher rates of cancellations?

5) What are the key factors influencing these cancellation rates?

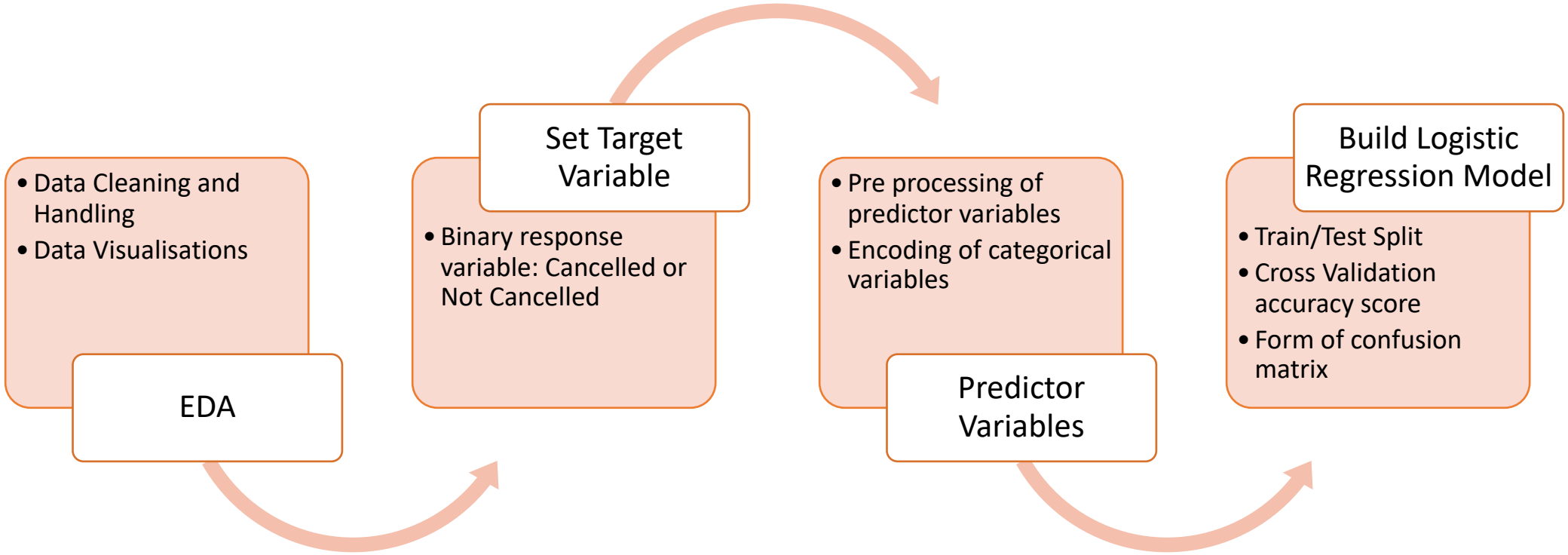
And many more questions can prop up whilst looking at the dataset.

Problem Statement

Predict the probabilities of cancellations to allow for hotel management to plan for future booking strategies.

- Aids hotel operators to better manage their resources by predicting the cancellation rates of their reservations.

Predictive Modelling



EDA

Variables

```
In [6]: index
```

```
Out[6]: RangeIndex(start=0, stop=119390, step=1)
```

```
In [7]: 
print ("rows and columns:",Hotel_bookings.shape)
```

```
rows and columns: (119390, 32)
```

```
In [8]: ("columns:",Hotel_bookings.columns)
```

```
Out[8]: ('columns:',
 Index(['hotel', 'is_canceled', 'lead_time', 'arrival_date_year',
        'arrival_date_month', 'arrival_date_week_number',
        'arrival_date_day_of_month', 'stays_in_weekend_nights',
        'stays_in_week_nights', 'adults', 'children', 'babies', 'meal',
        'country', 'market_segment', 'distribution_channel',
        'is_repeated_guest', 'previous_cancellations',
        'previous_bookings_not_canceled', 'reserved_room_type',
        'assigned_room_type', 'booking_changes', 'deposit_type', 'agent',
        'company', 'days_in_waiting_list', 'customer_type', 'adr',
        'required_car_parking_spaces', 'total_of_special_requests',
        'reservation_status', 'reservation_status_date'],
       dtype='object'))
```

Missing Values

```
In [12]: #null values in columns
         Hotel_bookings.isnull().sum()
```

```
Out[12]: 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                             0
         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
```

- Dataset contains 119390 rows.
- Out of which 112591 entries are missing for the company.
- Around 13% of the values under booking agent is missing.
- We note that 488 data points are missing for countries these guests come from.

Data Cleaning

- Hotel_bookings.country.fillna("Unknown", inplace=True)

```
In [16]: Hotel_bookings.isnull().sum()

Out[16]: 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                         0
         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
         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
```

Describe function

```
In [11]: Hotel_bookings.describe().T
```

Out[11]:

	count	mean	std	min	25%	50%	75%	max
is_canceled	119390.0	0.370416	0.482918	0.00	0.00	0.000	1.0	1.0
lead_time	119390.0	104.011416	106.863097	0.00	18.00	69.000	160.0	737.0
arrival_date_year	119390.0	2016.156554	0.707476	2015.00	2016.00	2016.000	2017.0	2017.0
arrival_date_week_number	119390.0	27.165173	13.605138	1.00	16.00	28.000	38.0	53.0
arrival_date_day_of_month	119390.0	15.798241	8.780829	1.00	8.00	16.000	23.0	31.0
stays_in_weekend_nights	119390.0	0.927599	0.998613	0.00	0.00	1.000	2.0	19.0
stays_in_week_nights	119390.0	2.500302	1.908286	0.00	1.00	2.000	3.0	50.0
adults	119390.0	1.856403	0.579261	0.00	2.00	2.000	2.0	55.0
children	119386.0	0.103890	0.398561	0.00	0.00	0.000	0.0	10.0
babies	119390.0	0.007949	0.097436	0.00	0.00	0.000	0.0	10.0
is_repeated_guest	119390.0	0.031912	0.175767	0.00	0.00	0.000	0.0	1.0
previous_cancellations	119390.0	0.087118	0.844336	0.00	0.00	0.000	0.0	26.0
previous_bookings_not_canceled	119390.0	0.137097	1.497437	0.00	0.00	0.000	0.0	72.0
booking_changes	119390.0	0.221124	0.652306	0.00	0.00	0.000	0.0	21.0
agent	103050.0	86.693382	110.774548	1.00	9.00	14.000	229.0	535.0
company	6797.0	189.266735	131.655015	6.00	62.00	179.000	270.0	543.0
days_in_waiting_list	119390.0	2.321149	17.594721	0.00	0.00	0.000	0.0	391.0
adr	119390.0	101.831122	50.535790	-6.38	69.29	94.575	126.0	5400.0
required_car_parking_spaces	119390.0	0.062518	0.245291	0.00	0.00	0.000	0.0	8.0
total_of_special_requests	119390.0	0.571363	0.792798	0.00	0.00	0.000	1.0	5.0

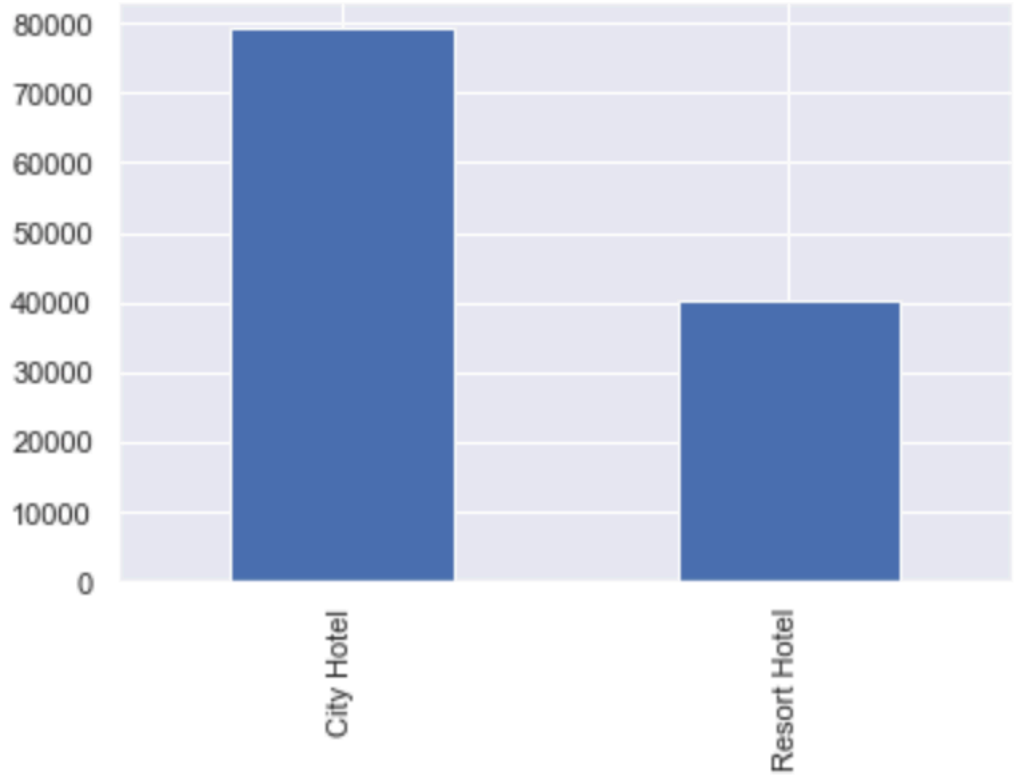
Numerical and Categorical Variables

```
In [19]: print_categories('hotel')
```

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

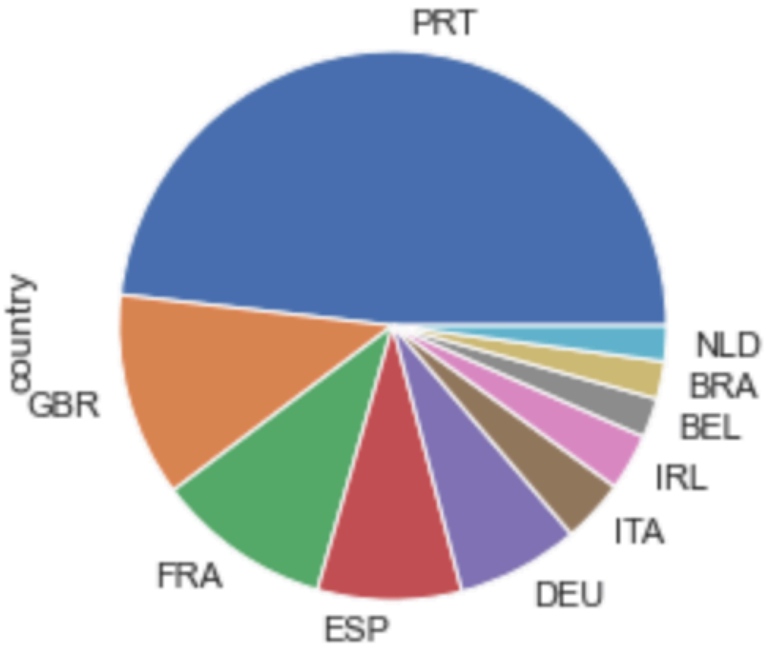
Hotel Bookings

- Resort Hotel : 40060 bookings
- City Hotel : 79330 bookings



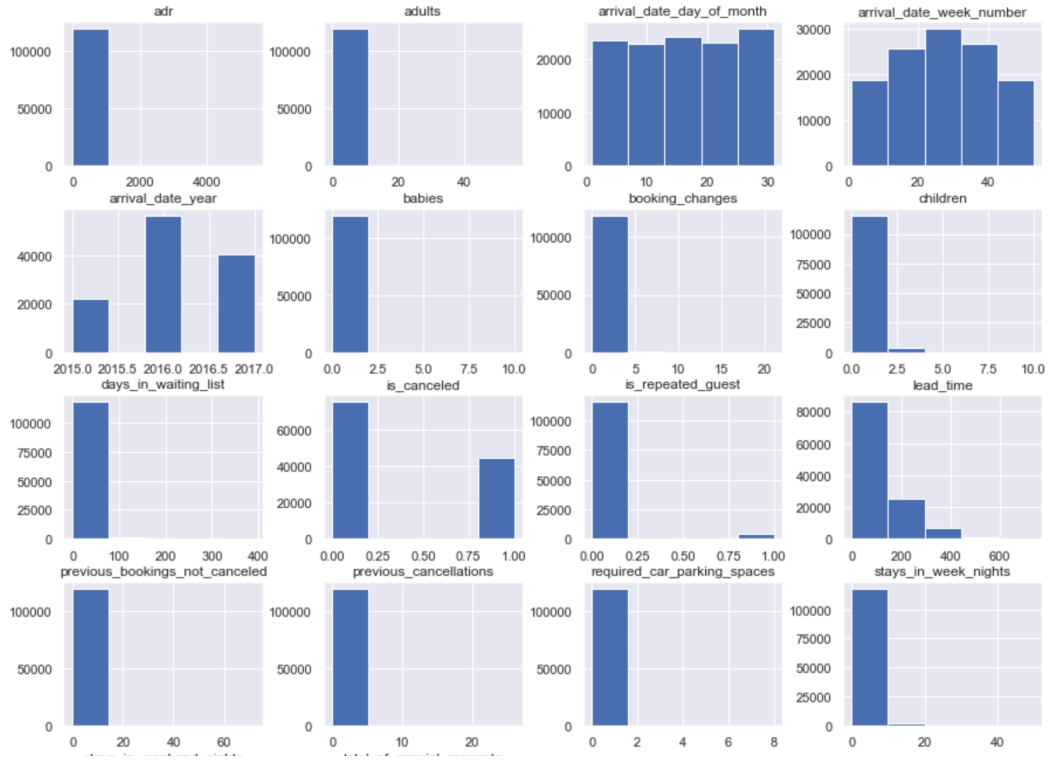
Where the guests are from?

- Hotel_bookings['country'].value_counts()[:10].plot(kind='pie');

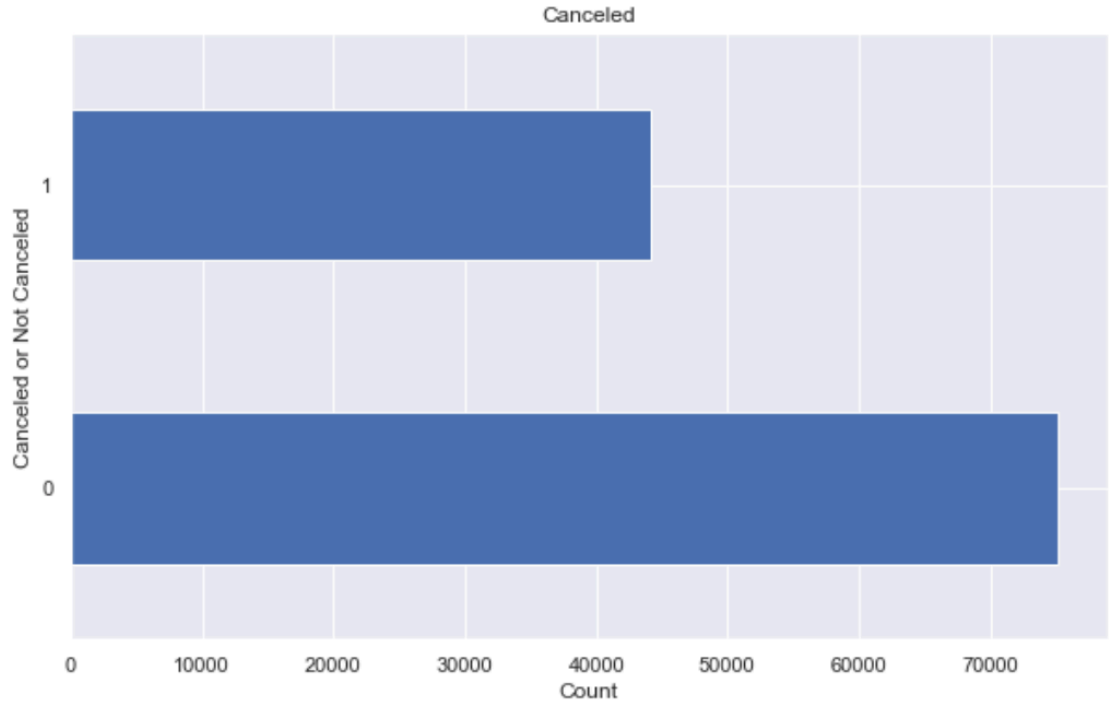


Distributions

- Hotel_bookings.hist(bins=5,figsize=(15, 14))
- plt.show()



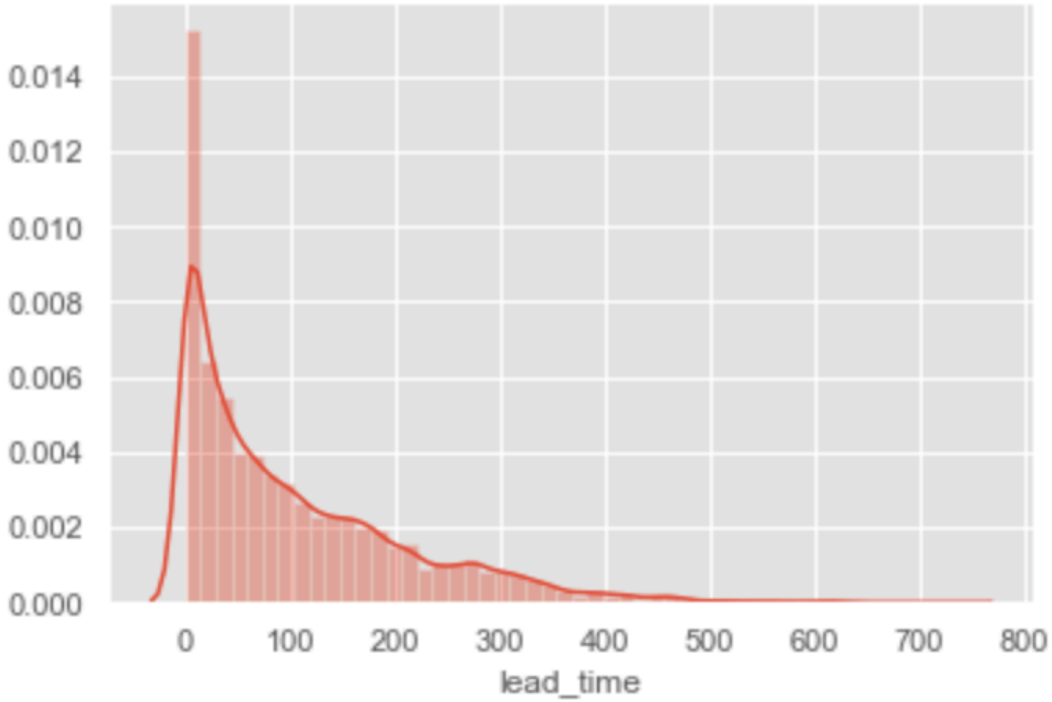
Cancellation rate

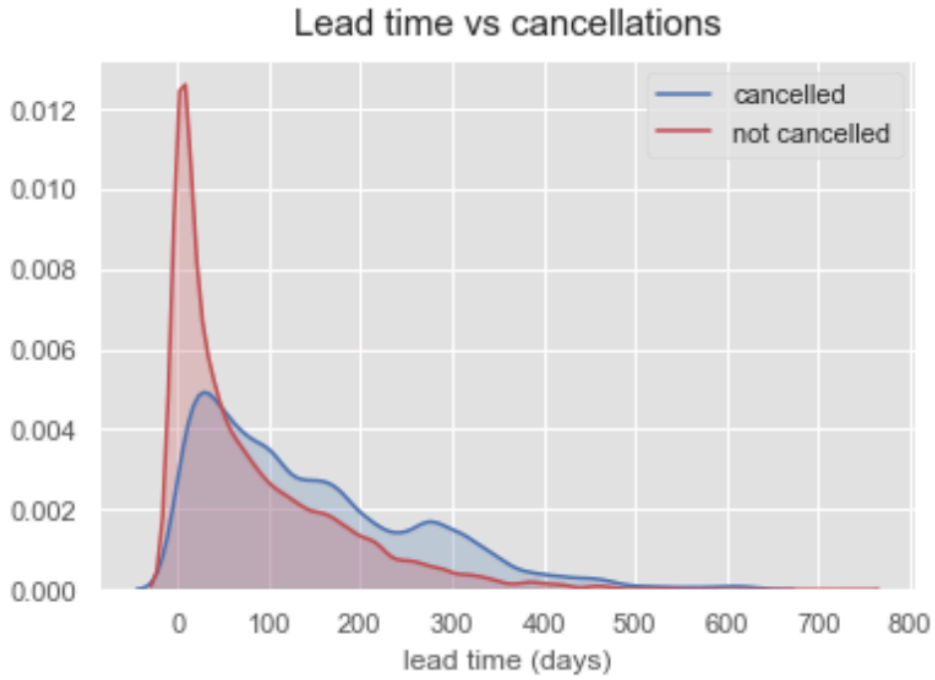


- Around 34% of bookings were cancelled.

Lead Time

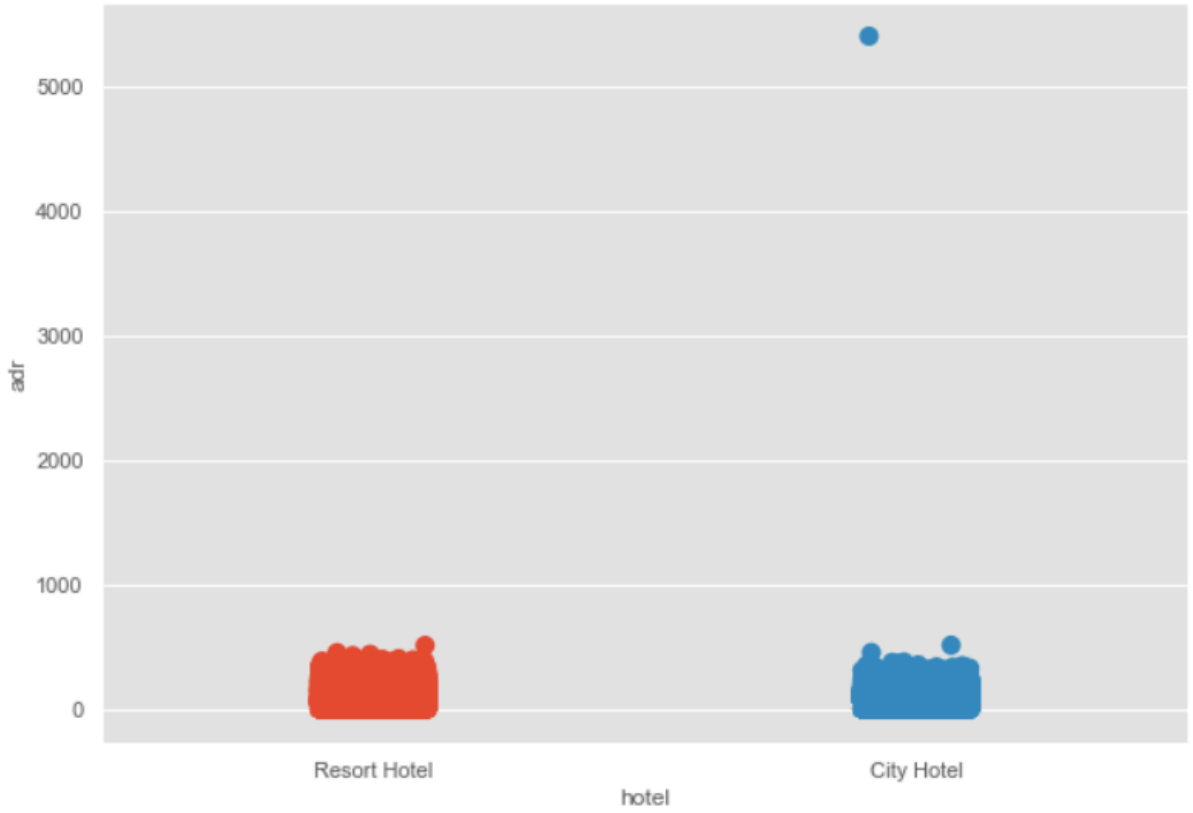
How in advance did hotel guests booked and relation with cancellation?





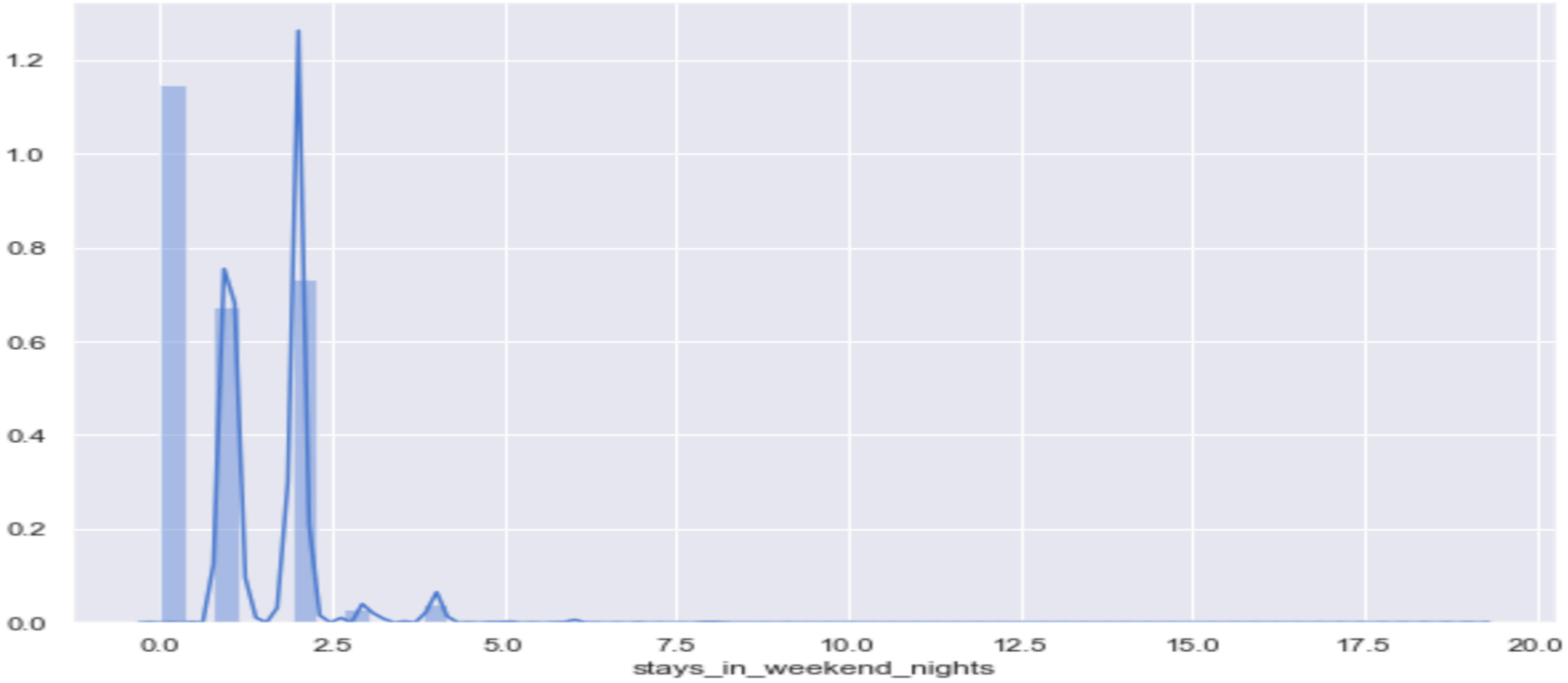
Average Daily Rate

- Average Daily Rate (ADR)= Rooms Revenue Earned/ Number of Rooms Sold



Stay Weekend Nights

How many of these bookings were for weekend night stays?

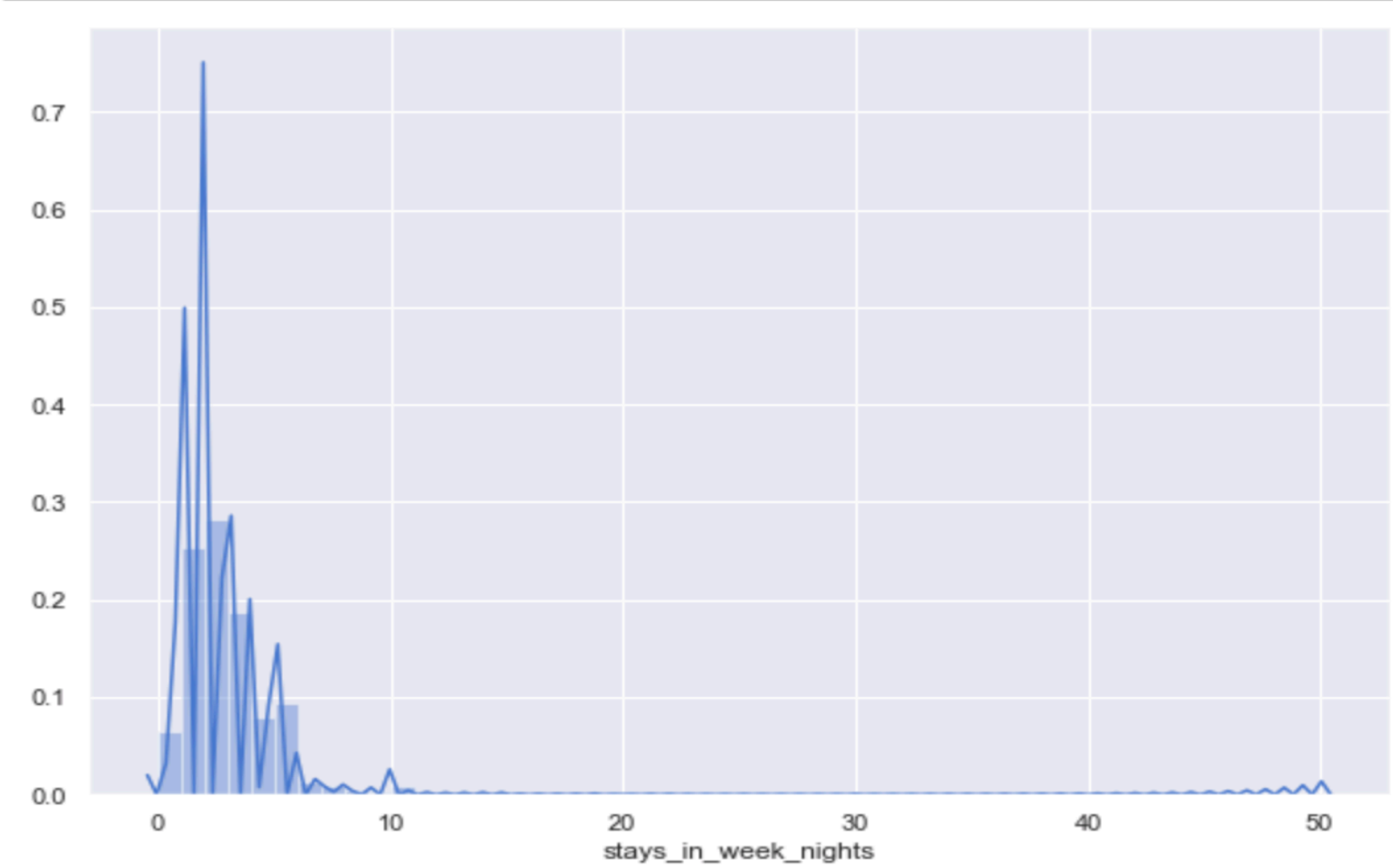


Stay Weekend Nights

How many of these bookings were for weekend night stays and out of these bookings, what was the cancellation rate?

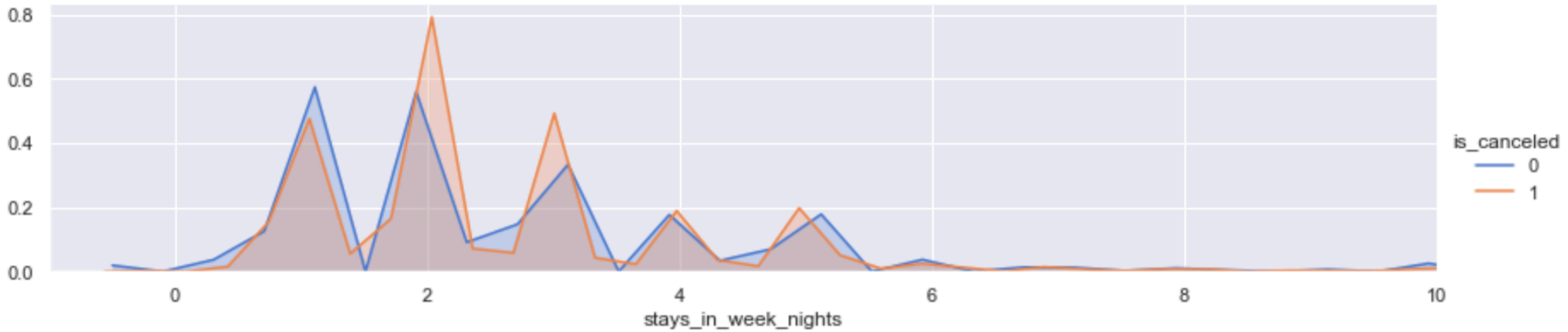
Stay Weekday Nights

How many of these bookings were for weekend night stays?



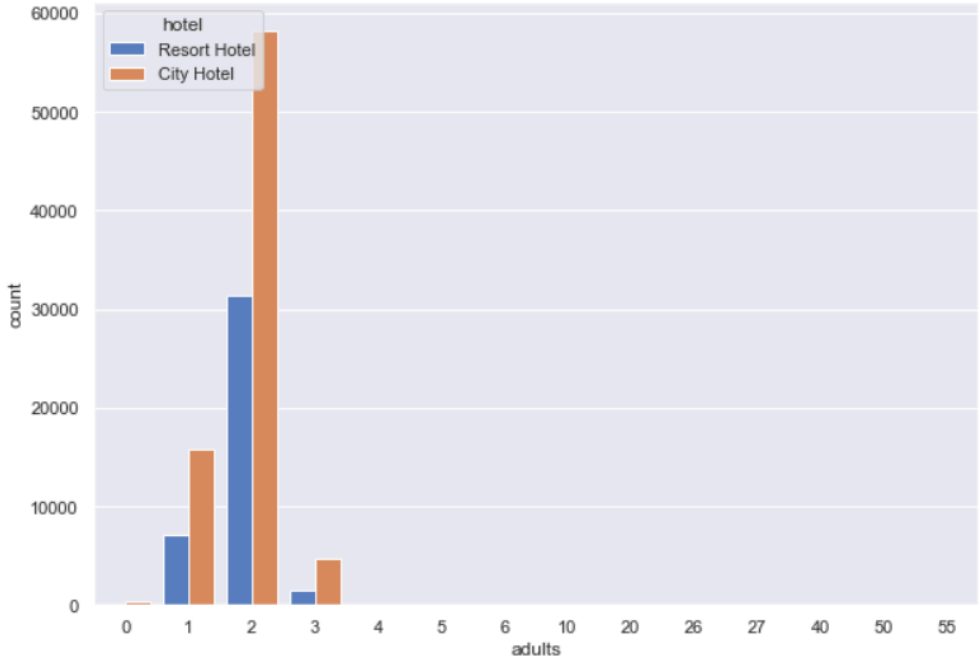
Stay Weekday Nights

How many of these bookings were for weekend night stays and out of these bookings, what was the cancellation rate?



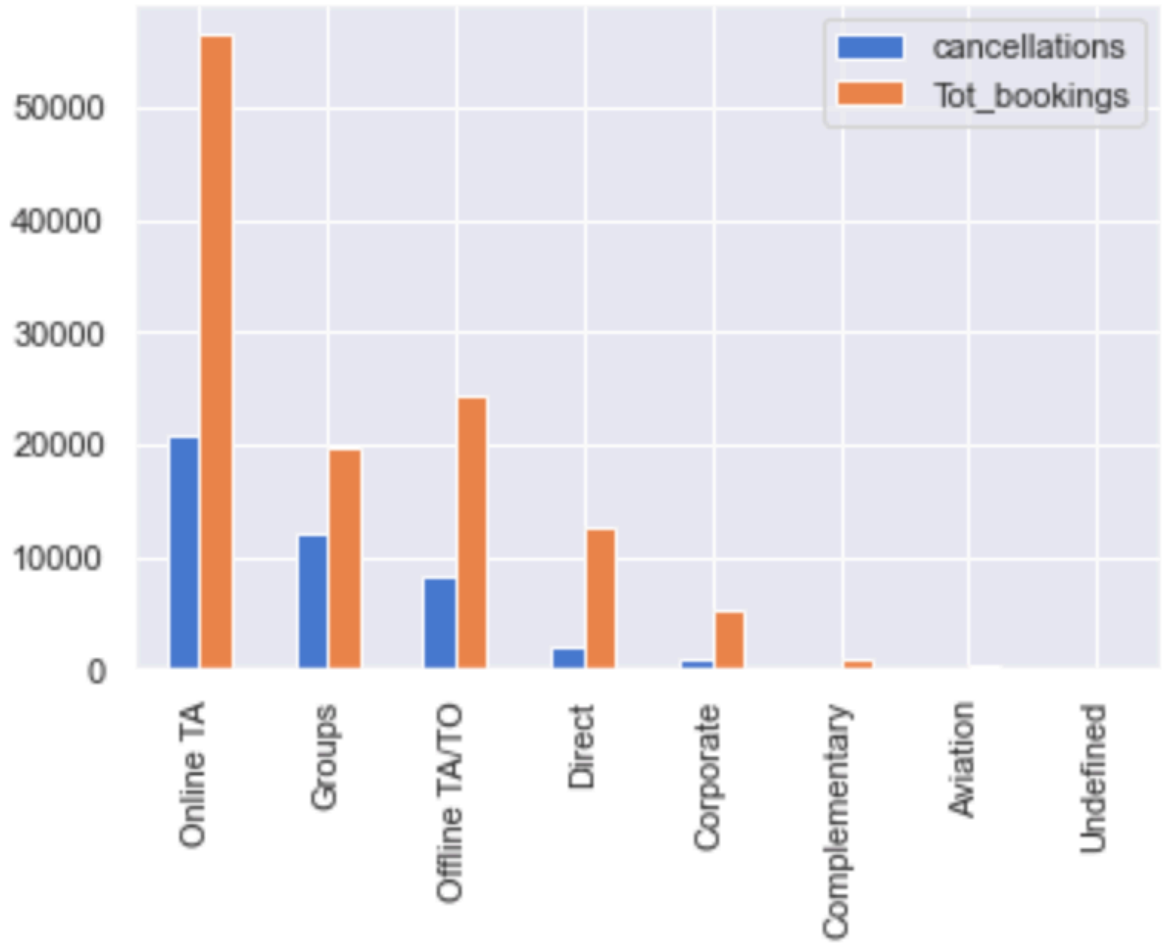
Types of guests

Distributions of adults and children guests?



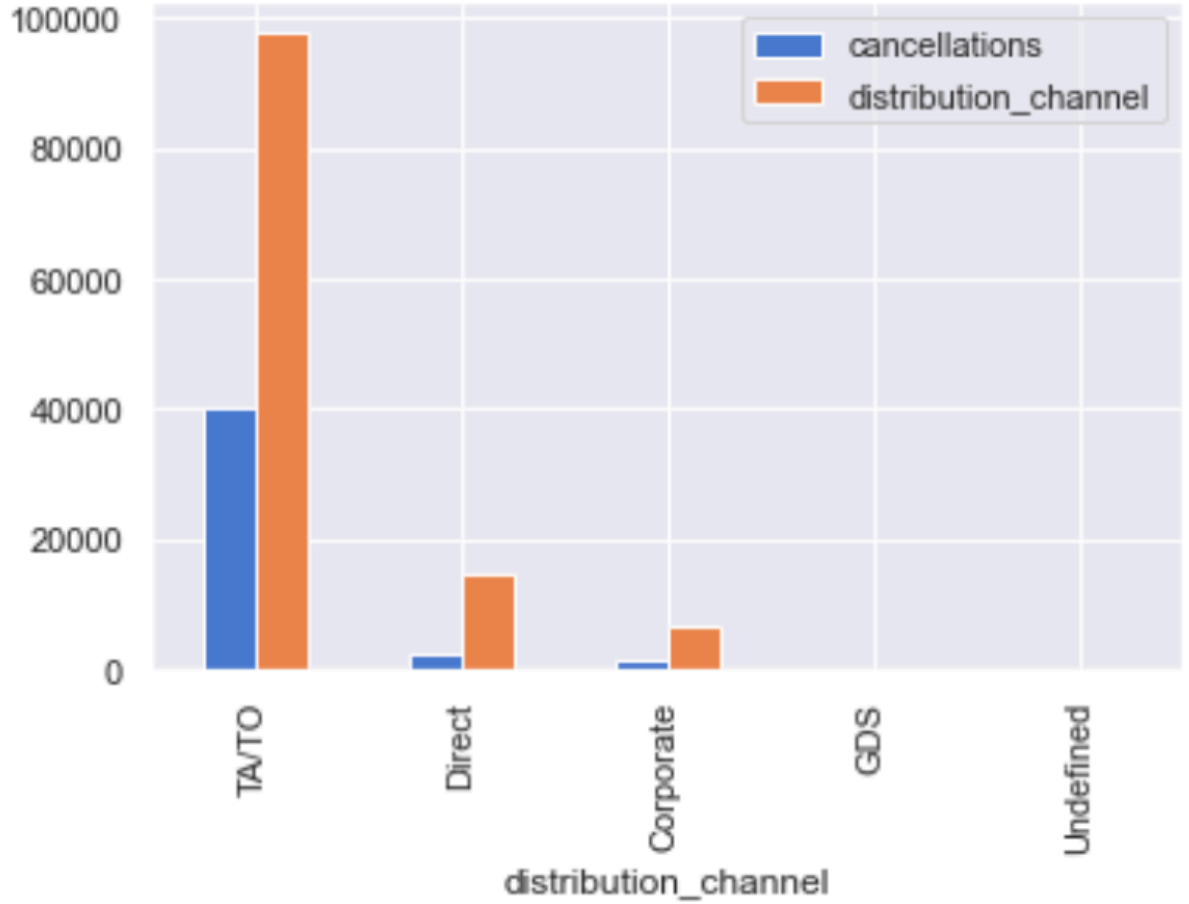
Market Segments

How these guests booked their rooms and out of which what's the cancellation rate?



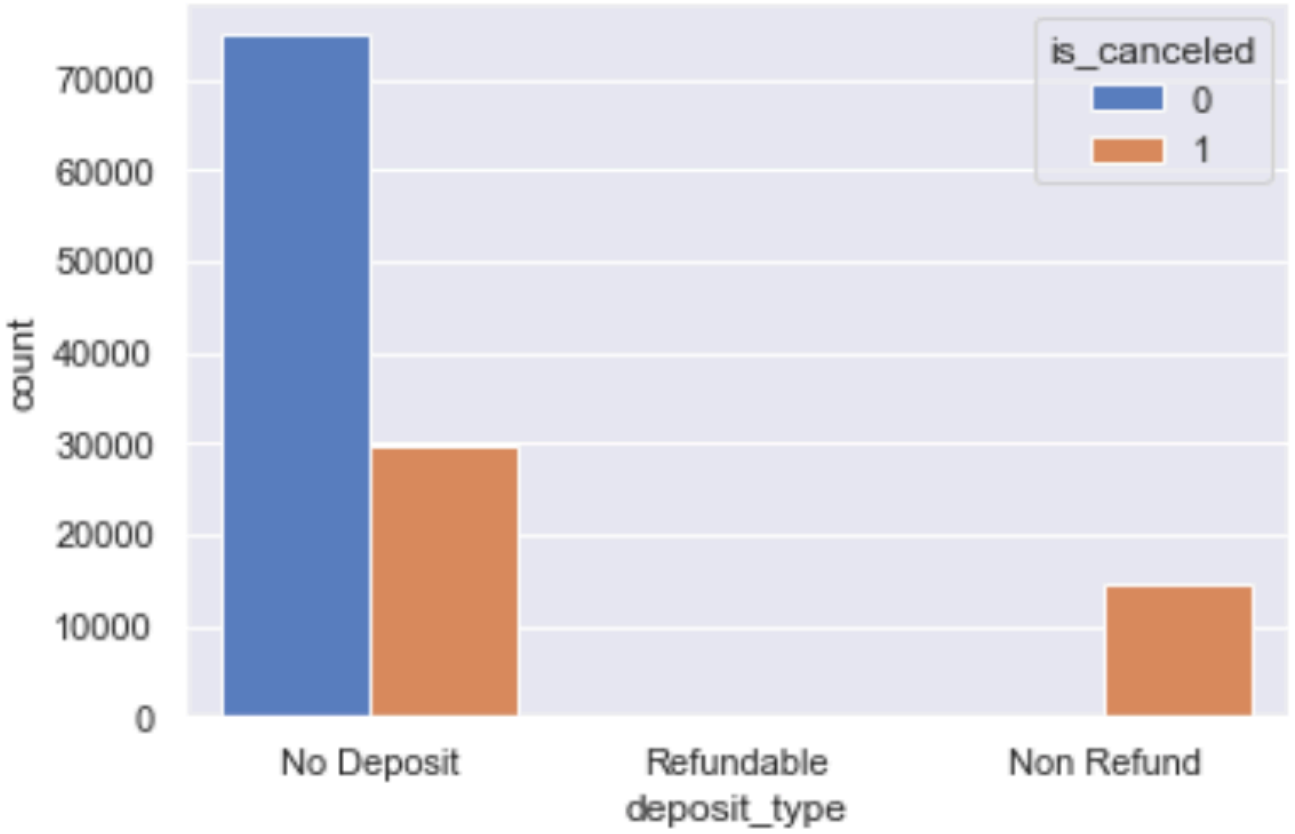
Distribution Channels

How both hotels promoted and sold rooms and out of which what's the cancellation rate?



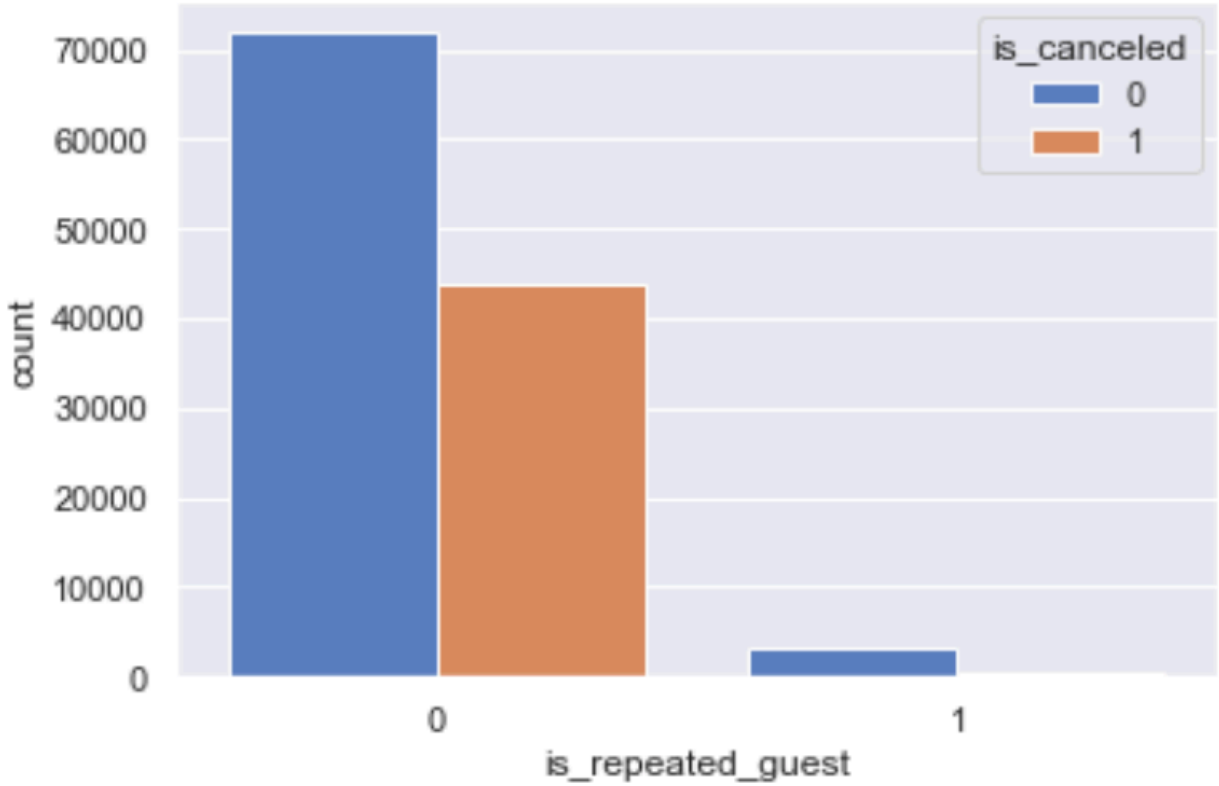
Deposit Type

How many of the guests paid deposits/no deposits and what was the cancellation like?



Repeated guest

How many of the guests are repeated and does this mean lesser cancellation from these groups?

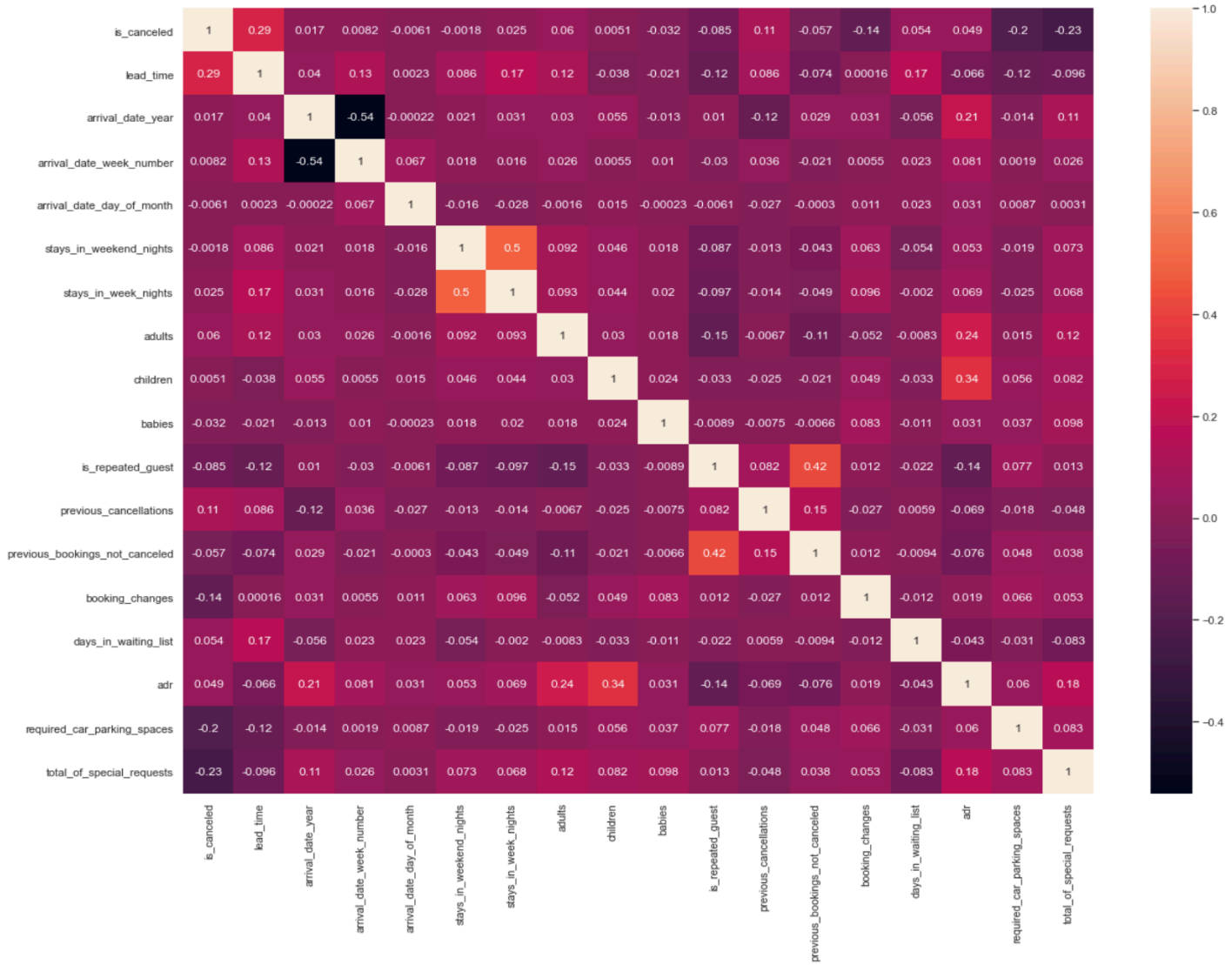


Correlation Matrix

- corr_matrix = Hotel_bookings.corr()
- print(corr_matrix["is_canceled"].sort_values(ascending=False))

```
is_canceled                      1.000000
lead_time                        0.293133
previous_cancellations           0.110135
adults                           0.060015
days_in_waiting_list             0.054188
adr                              0.048708
stays_in_week_nights             0.024773
arrival_date_year                0.016662
arrival_date_week_number         0.008160
children                         0.005051
stays_in_weekend_nights         -0.001781
arrival_date_day_of_month       -0.006142
babies                          -0.032490
previous_bookings_not_canceled  -0.057357
is_repeated_guest               -0.084792
booking_changes                 -0.144396
required_car_parking_spaces     -0.195496
total_of_special_requests       -0.234652
Name: is_canceled, dtype: float64
```

Heatmap



- Note the correlation is highest between lead time and canceled variable.
- Even then, relatively low correlation value of 0.29.

Set target variable

Target Variable

We aim predict the cancellation rate among the hotel guests. Hence, our target variable will be set is "is_canceled".

In [76]:
```
#is_cancelled is the target variable.
Hotel_bookings['is_canceled']
```

Out[76]:
```
0         0
1         0
2         0
3         0
4         0
         ..
119385    0
119386    0
119387    0
119388    0
119389    0
Name: is_canceled, Length: 119389, dtype: int64
```

In [77]:
```
y= Hotel_bookings['is_canceled']
```

Predictor variables

Set Predictor Variables

Pre processing of predictor variables

- Drop reservation status column as directly related to Target variable (Checked out or Cancelled).
- Hotel_bookings = Hotel_bookings.drop (columns= ['reservation_status', 'all_children','adults','children','babies','reservation_status_date'])

- Create a variable of “total_guests”

```
In [79]: Hotel_bookings['total_guests'] =Hotel_bookings['adults'] + Hotel_bookings['children'] + Hotel_bookings['babies']

In [80]: Hotel_bookings['total_guests'].value_counts()

Out[80]: 2.0     82047
         1.0     22581
         3.0     10494
         4.0      3929
         0.0       180
         5.0       137
         26.0        5
         27.0        2
         12.0        2
         10.0        2
         20.0        2
         55.0        1
         6.0         1
         50.0        1
         40.0        1
         Name: total_guests, dtype: int64

In [81]: Hotel_bookings.drop(Hotel_bookings.loc[Hotel_bookings['total_guests']==0].index, inplace=True)

In [82]: Hotel_bookings['total_guests'].value_counts()

Out[82]: 2.0     82047
         1.0     22581
         3.0     10494
         4.0      3929
         5.0       137
         26.0        5
         27.0        2
         12.0        2
         10.0        2
         20.0        2
         55.0        1
         6.0         1
         50.0        1
         40.0        1
         Name: total_guests, dtype: int64
```

Dummy Variables

Encoding the nominal categorical variables.

```
#Encode categorical variables to run logistic regression model

Hotel_bookings_with_dummies = pd.get_dummies(data = Hotel_bookings, columns = ['hotel','arrival_date_month', 'arriv
                  'distribution_channel', 'reserved_room_type', 'assigned_room_type', 'deposit_type', 'customer_ty
                                        prefix = ['hotel','arrival month', 'arrival year','meal','market segment',
                  'distribution channel', 'reserved room', 'assigned room', 'deposit', 'customer', 'country'] )
Hotel_bookings_with_dummies.head()
```

	is_canceled	lead_time	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	is_repeated_guest	previous_canc
0	0	342	27	1	0	0	0	
1	0	737	27	1	0	0	0	
2	0	7	27	1	0	1	0	
3	0	13	27	1	0	1	0	
4	0	14	27	1	0	2	0	

5 rows × 255 columns

Set Predictor Variables

Take all columns except target as predictor columns (254 columns)

In [91]:
```
# Take all columns except target as predictor columns
predictor_columns = [c for c in Hotel_bookings_with_dummies.columns if c != 'is_canceled']
# Load the dataset as a pandas data frame
X = pd.DataFrame(Hotel_bookings_with_dummies, columns = predictor_columns)
```

In [92]:
```
X
```

Out[92]:

	lead_time	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	is_repeated_guest	previous_cancellations
0	342	27	1	0	0	0	0
1	737	27	1	0	0	0	0
2	7	27	1	0	1	0	0
3	13	27	1	0	1	0	0
4	14	27	1	0	2	0	0
...	...	...	...	...	...	...	...
119385	23	35	30	2	5	0	0
119386	102	35	31	2	5	0	0
119387	34	35	31	2	5	0	0
119388	109	35	31	2	5	0	0
119389	205	35	29	2	7	0	0

119209 rows × 254 columns

Build logistic regression model

Create & Fit Model

In [93]:
```
#Target variable
y= Hotel_bookings['is_canceled']
```

In [94]:
```
from sklearn.linear_model import LogisticRegression
logreg=LogisticRegression()
```

In [95]:
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)
```

Cross Validation Accuracy Score

- Cross validation accuracy score is more accurate estimate of out-of-sample accuracy.

In [97]:
```
from sklearn.model_selection import cross_val_score
scores = cross_val_score(logreg, X_train, y_train, cv=5)
print('Cross-Validation Accuracy Scores', scores)
```

- Score for each fold:
- Cross-Validation Accuracy Scores [0.80643808 0.80554682 0.80936402 0.80527447 0.79961202]

Cross Validation Accuracy Score

- Cross validation mean score:

```
In [98]: scores = pd.Series(scores)
         scores.min(), scores.mean(), scores.max()

Out[98]: (0.7996120169873643, 0.8052470824212141, 0.8093640224400985)
```

- Score for each fold:
- Score allows us to see how well our model generalizes and accuracy score is pretty good at 0.80.

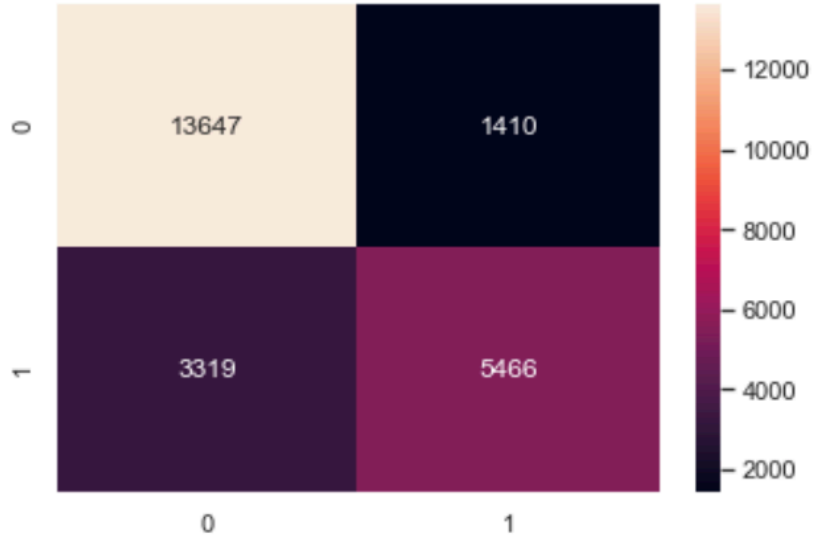
Confusion Matrix

```
In [99]: from sklearn import metrics

In [100]: #confusion matrix
          y_pred = logreg.predict (X_test)
          print(metrics.confusion_matrix(y_test,y_pred))

          [[13647  1410]
           [ 3319  5466]]
```

- True Positives (TP): we correctly predicted the room cancellations 5466
- True Negatives (TN): we correctly predicted the non cancellations 13647.
- False Positives (FP): we incorrectly predicted the room cancellations 1410.
- False Negatives (FN): we incorrectly predicted that the non cancellations 3319.

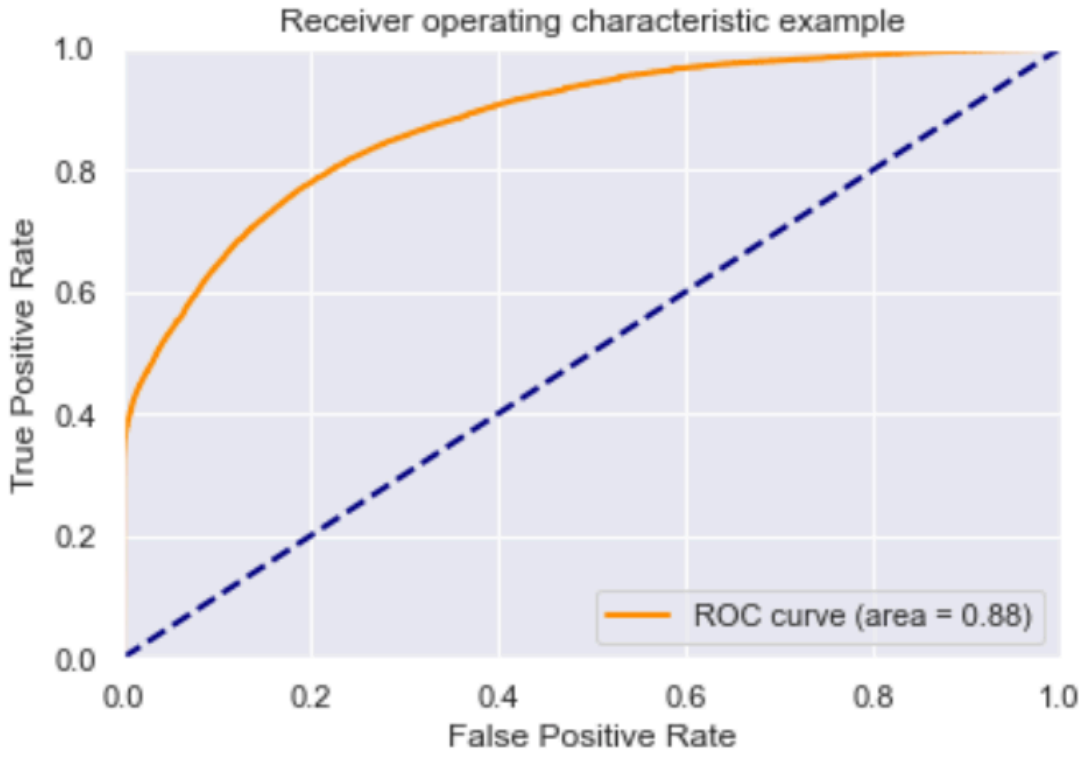


Classification Report

	precision	recall	f1-score	support
0	0.80	0.91	0.85	15057
1	0.79	0.62	0.70	8785
accuracy			0.80	23842
macro avg	0.80	0.76	0.78	23842
weighted avg	0.80	0.80	0.80	23842

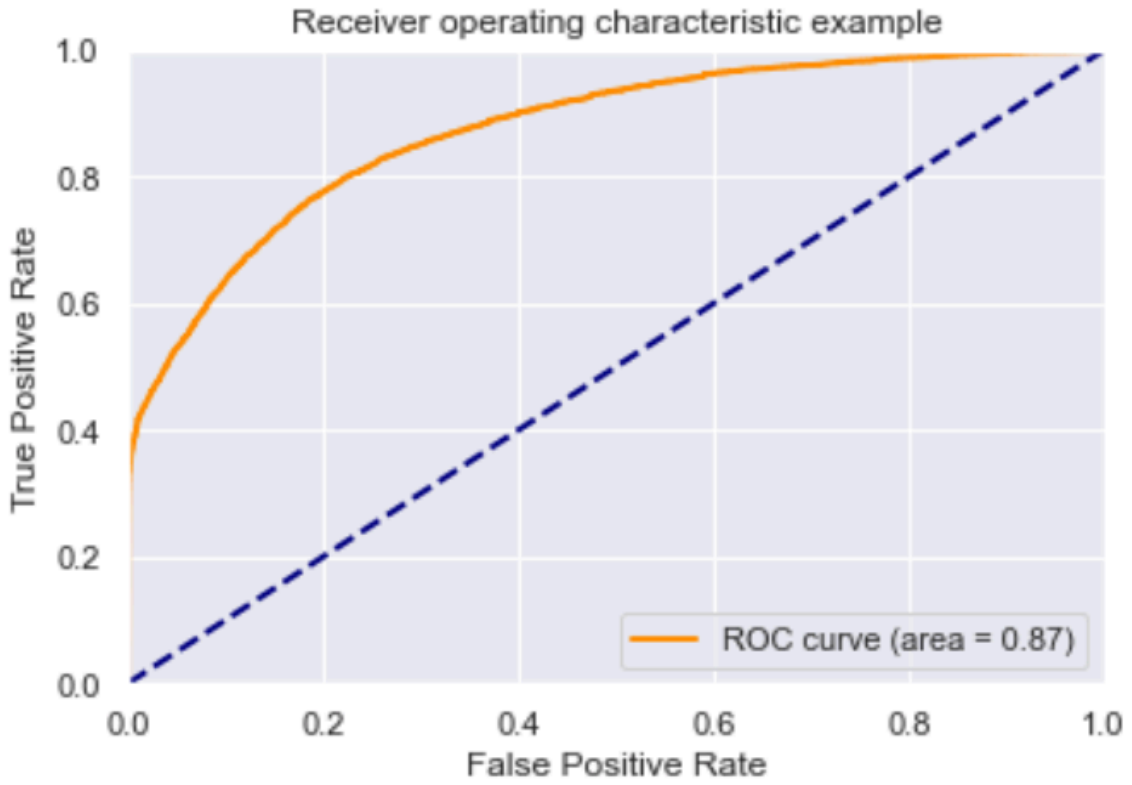
ROC Curve

(Train Set)



ROC Curve

(Test Set)



Results

Evaluation of Results

- In this prediction case, when our model predicted room booking cancellations, 80% of the time there has been cancellations (predictions were correct).
- With 80% accuracy, more than eight times out of ten, our model correctly predicts whether a reservation will be successfully booked or cancelled.
- AUC score for the case is 0.88 for the train and 0.87 test sets, representing a good classifier. Overfitting is unlikely.

References

- https://analytx4t.com/predicting-travel-booking-cancellations-with-machine-learning/
- https://www.kaggle.com/wirachleelakiatiwong/it-will-be-canceled-or-not-eda-and-modeling
- https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8
- https://www.kaggle.com/rgoyalml/hotel-booking-demand-eda-ml
- https://www.ritchieng.com/machine-learning-evaluate-classification-model/
- https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python


Thank you