

Final Project

Bramanti coneri

Bramanti Coneri Team Member



**ADE RIZKY
RAHMANIA**



BIJA HARDHONA AJI



REZA ALFADIN



**SISKA WULANDHARI
BIANPUTRI**



WIWIT WIDIYANTI

TABLE OF CONTENT



Data Summary



resampling to handle imbalanced dataset



Problem Statements



Kesimpulan



Identify



Data Prep-Processing



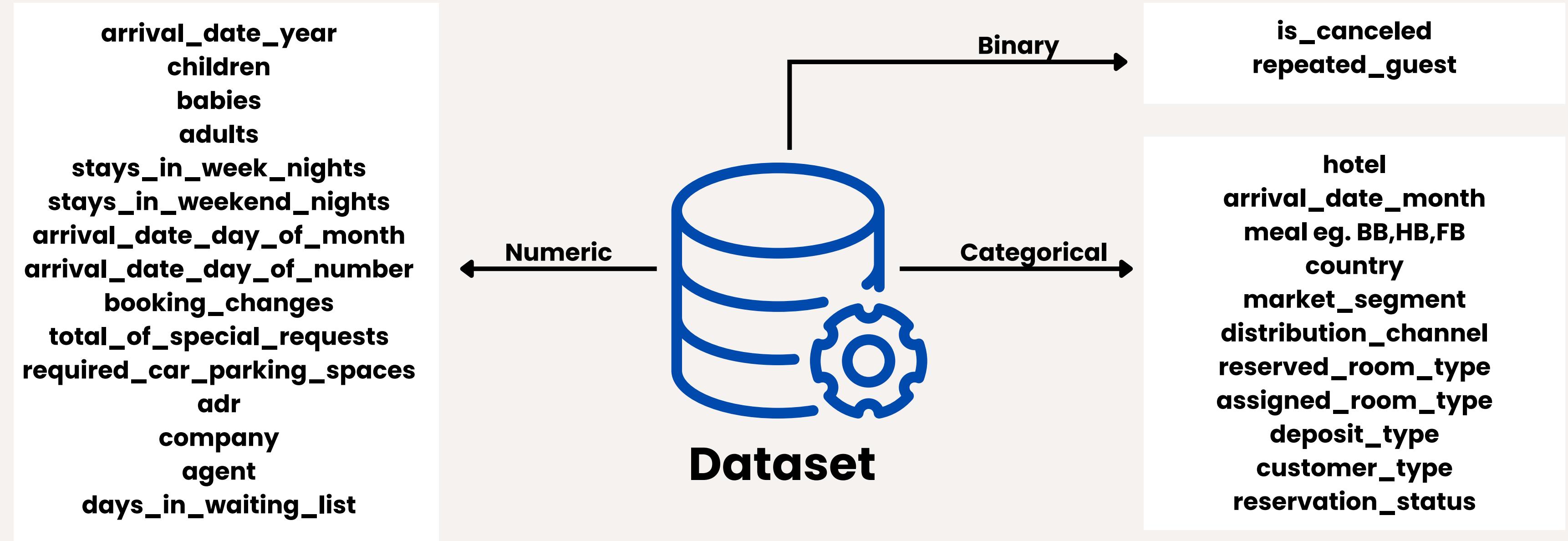
Modelling

Data Summary

Hotel booking demand dataset ini menggambarkan dua dataset dengan data permintaan hotel. Salah satu hotel (H1) adalah hotel resor dan yang lainnya adalah hotel kota (H2). Kedua kumpulan data memiliki struktur yang sama, dengan 32 variabel yang menggambarkan 40.060 pengamatan H1 dan 79.330 pengamatan H2. Setiap pengamatan mewakili pemesanan hotel. Kedua dataset tersebut adalah data pemesanan yang akan tiba antara tanggal 1 Juli 2015 hingga 31 Agustus 2017 termasuk pemesanan yang tiba secara efektif dan pemesanan yang dibatalkan.



Data Summary



Data Summary

hotel	Hotel (H1 = Resort Hotel atau H2 = City Hotel)
is_canceled	Nilai yang menunjukkan apakah pemesanan dibatalkan (1) atau tidak (0)
lead_time	Jumlah hari yang berlalu antara tanggal masuk pemesanan ke dalam (Property Management System) PMS dan tanggal kedatangan
arrival_date_year	Tahun kedatangan tanggal
arrival_date_month	Bulan tanggal kedatangan
arrival_date_week_number	Minggu keberapa dalam tahun untuk tanggal kedatangan
arrival_date_day_of_month	Hari tanggal kedatangan
stays_in_weekend_nights	Jumlah malam akhir pekan (Sabtu atau Minggu) tamu menginap atau memesan untuk menginap di hotel
stays_in_week_nights	Jumlah malam minggu (Senin sampai Jumat) tamu menginap atau memesan untuk menginap di hotel
adults	Jumlah orang dewasa
children	Jumlah anak
babies	Jumlah bayi
meal	Jenis makanan yang dipesan. Kategori yang disajikan dalam paket makanan standar perhotelan: Tidak ditentukan/SC - tidak ada paket makan; BB - Tempat Tidur & Sarapan; HB - Half board (sarapan dan satu kali makan lainnya - biasanya makan malam); FB - Full board (sarapan, makan siang, dan makan malam)

Data Summary

country object	Negara Asal. Kategori diwakili dalam format ISO 3155-3:2013
market_segment	Segmen pasar. Dalam kategori, istilah
TA - Agen Perjalanan	
TO - Operator Tur	
distribution_channel	Saluran distribusi pemesanan. Dalam kategori, istilah
TA - Agen Perjalanan	
TO - Operator Tur	
is_repeated_guest	Nilai yang menunjukkan apakah nama pemesanan berasal dari tamu yang pernah berkunjung sebelumnya (1) atau bukan (0)
previous_cancellations	Jumlah pemesanan sebelumnya yang dibatalkan oleh pelanggan sebelum pemesanan saat ini
previous_bookings_not_canceled	Jumlah pemesanan sebelumnya yang tidak dibatalkan oleh pelanggan sebelum pemesanan saat ini
reserved_room_type	Kode tipe kamar yang dipesan.
assigned_room_type	Kode untuk jenis kamar yang ditetapkan untuk pemesanan. Terkadang jenis kamar yang ditetapkan berbeda dari jenis kamar yang dipesan karena alasan operasi hotel (misalnya pemesanan berlebih) atau karena permintaan pelanggan.
booking_changes :	Jumlah perubahan/perubahan yang dilakukan pada pemesanan dari saat pemesanan dimasukkan pada PMS hingga saat check-in atau pembatalan.
deposit_type :	Indikasi jika pelanggan melakukan deposit untuk menjamin pemesanan. Variabel ini dapat mengasumsikan tiga kategori: No Deposit – tidak ada deposit yang dilakukan; Non Refund – deposit dilakukan sebesar total biaya menginap; Refundable - deposit dilakukan dengan nilai di bawah total biaya menginap

Kesimpulan Data

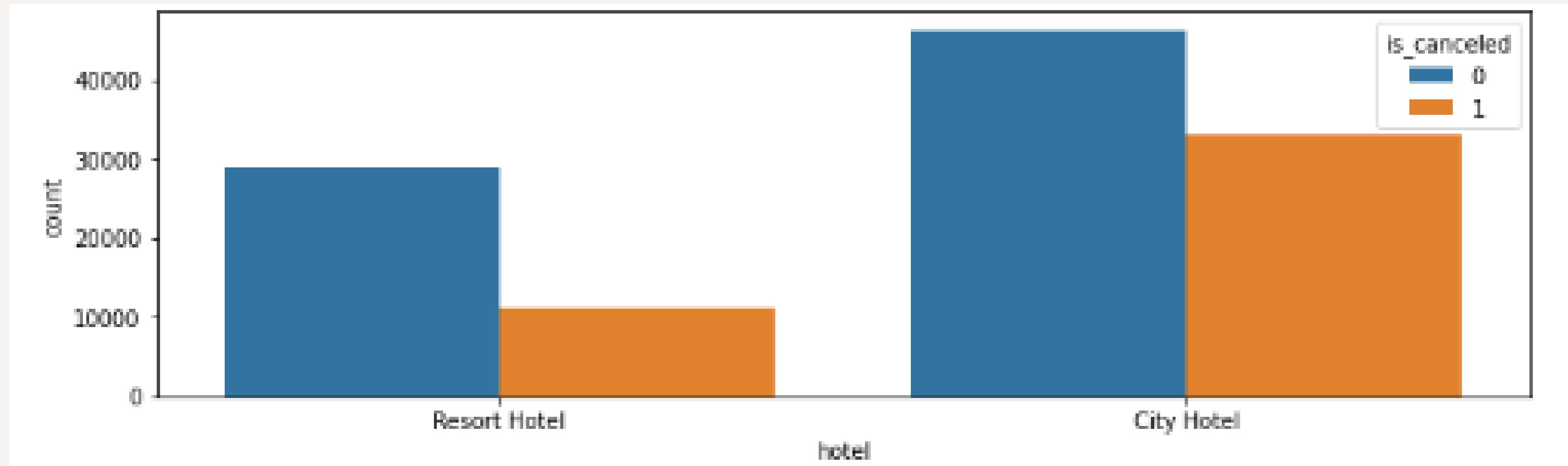
agent	ID agen perjalanan yang melakukan pemesanan
company	ID perusahaan/entitas yang melakukan pemesanan atau bertanggung jawab untuk membayar pemesanan.
days_in_waiting_list	Jumlah hari pemesanan dalam daftar tunggu sebelum dikonfirmasi ke pelanggan
customer_type	Jenis pemesanan, dengan asumsi salah satu dari empat kategori: Contract - ketika pemesanan memiliki penjatahan atau jenis kontrak lain yang terkait dengannya; Group - saat pemesanan dikaitkan dengan grup; Transient - ketika pemesanan bukan bagian dari grup atau kontrak, dan tidak terkait dengan pemesanan sementara lainnya; Transient-party - saat pemesanan bersifat sementara, tetapi terkait dengan setidaknya pemesanan sementara lainnya
adr	Average Daily Rate atau Tarif Harian Rata-Rata sebagaimana didefinisikan dengan membagi jumlah semua transaksi penginapan dengan jumlah total malam menginap
required_car_parking_spaces	Jumlah tempat parkir mobil yang dibutuhkan oleh pelanggan
total_of_special_requests	Jumlah permintaan khusus yang dibuat oleh pelanggan (misalnya tempat tidur kembar atau lantai atas)
reservation_status	Status terakhir reservasi, dengan asumsi salah satu dari tiga kategori: Canceled - pemesanan dibatalkan oleh pelanggan; Check-Out - pelanggan telah check-in tetapi sudah berangkat; No-Show - pelanggan tidak check-in dan memberi tahu hotel tentang alasannya
reservation_status_date	Tanggal saat status terakhir ditetapkan. Variabel ini dapat digunakan bersama dengan ReservationStatus untuk mengetahui kapan pemesanan dibatalkan atau kapan pelanggan check-out dari hotel

Problem Statements

Bagaimana melakukan prediksi customer yang canceled berdasarkan PMS (Property Management System) database hotel.

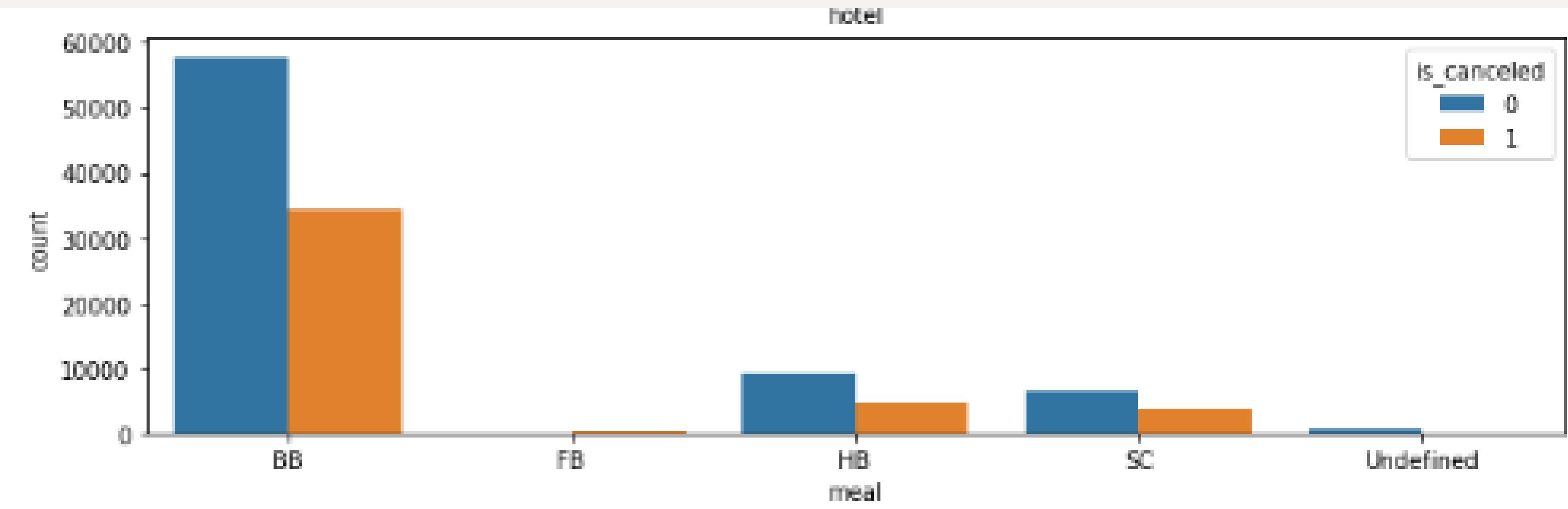
IDENTIFY

HOTEL



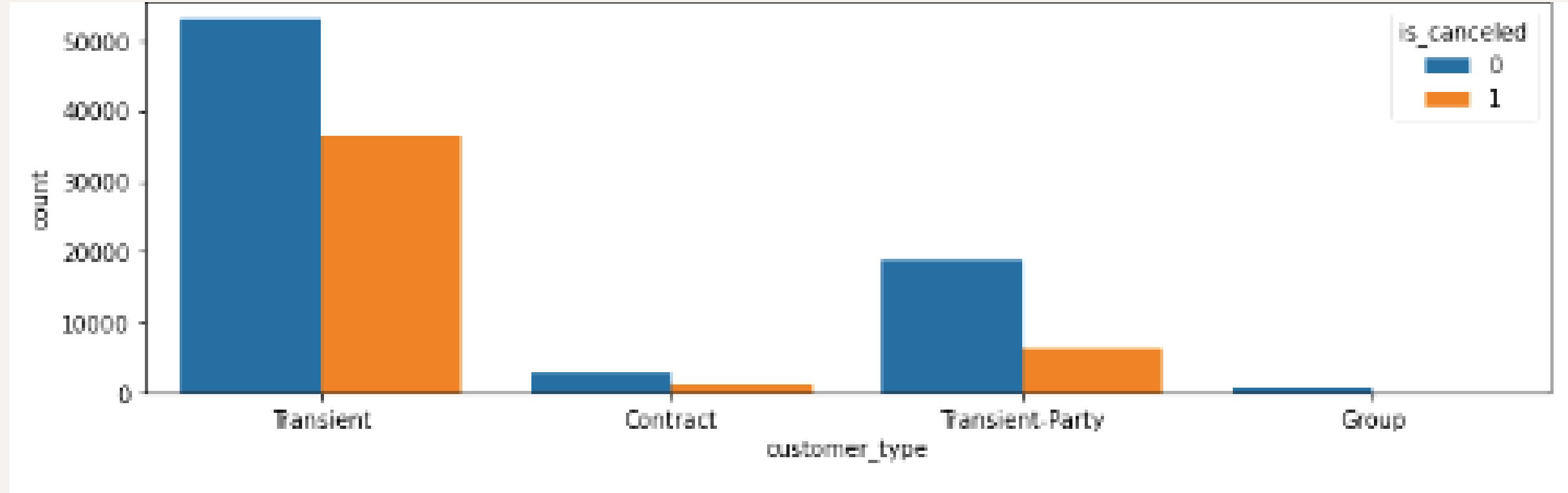
Mayoritas customer memilih dan terjadi canceled
yaitu pada city hotel .

MEAL



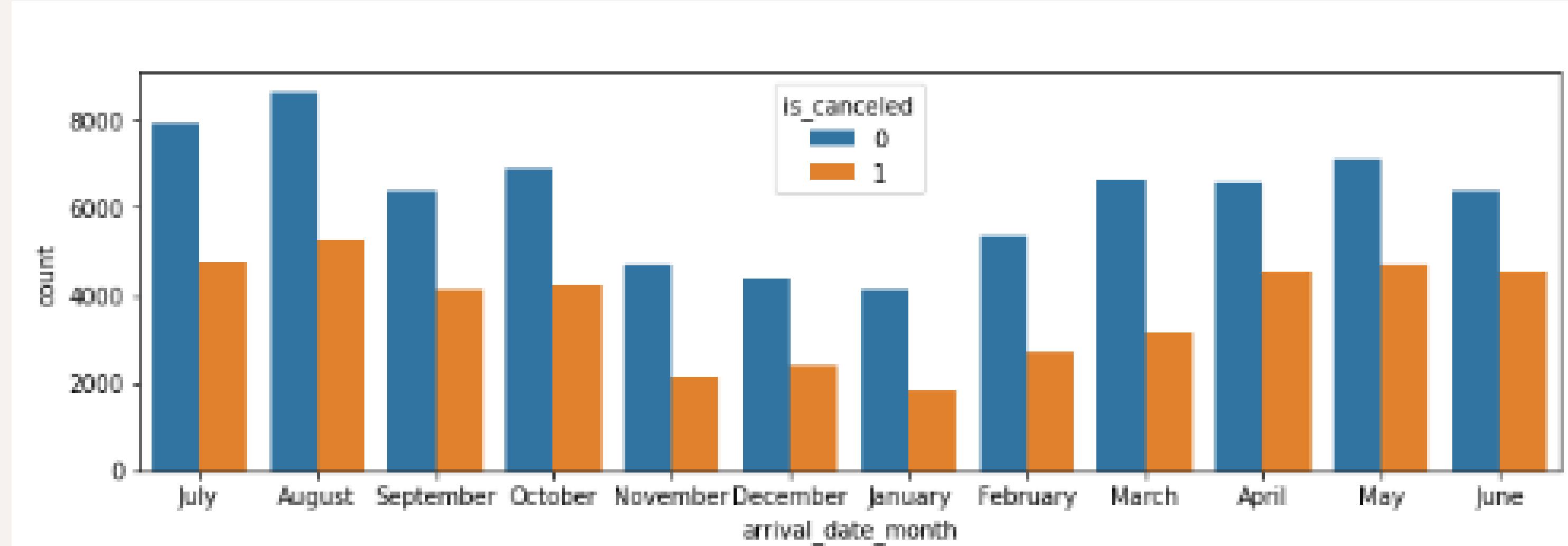
Mayoritas customer memilih paket meal BB atau Bed&Breakfast namun canceled terbanyak berdasarkan perbandingan ialah HB

Customer type



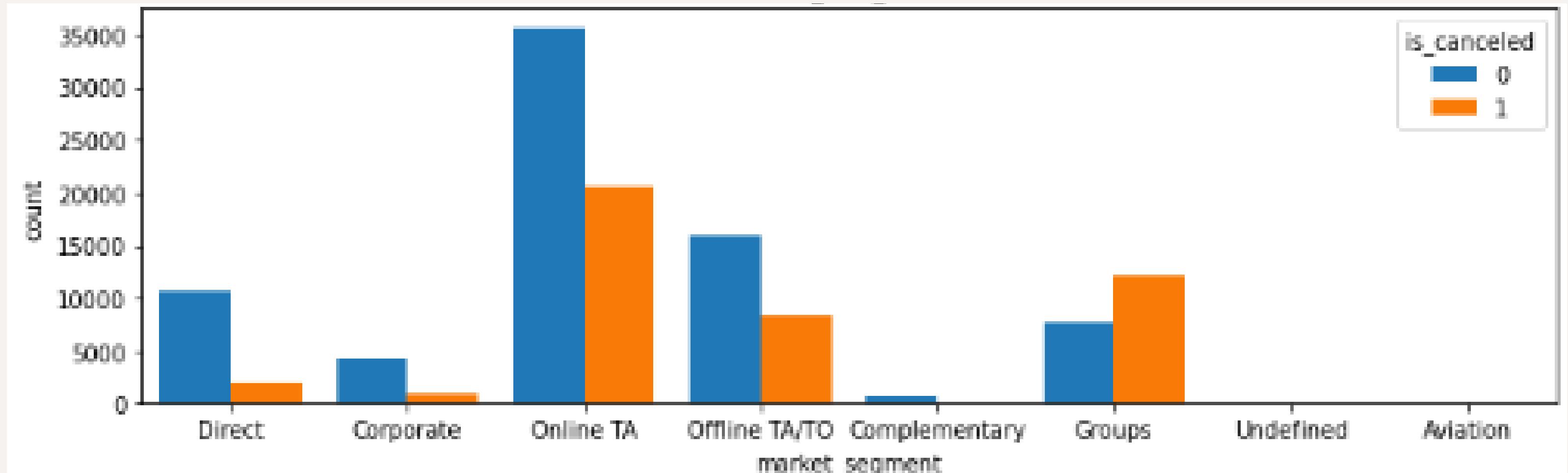
Customer type terbanyak pada transient dan pada grup tidak ada customer yang melakukan canceled

arrive date month



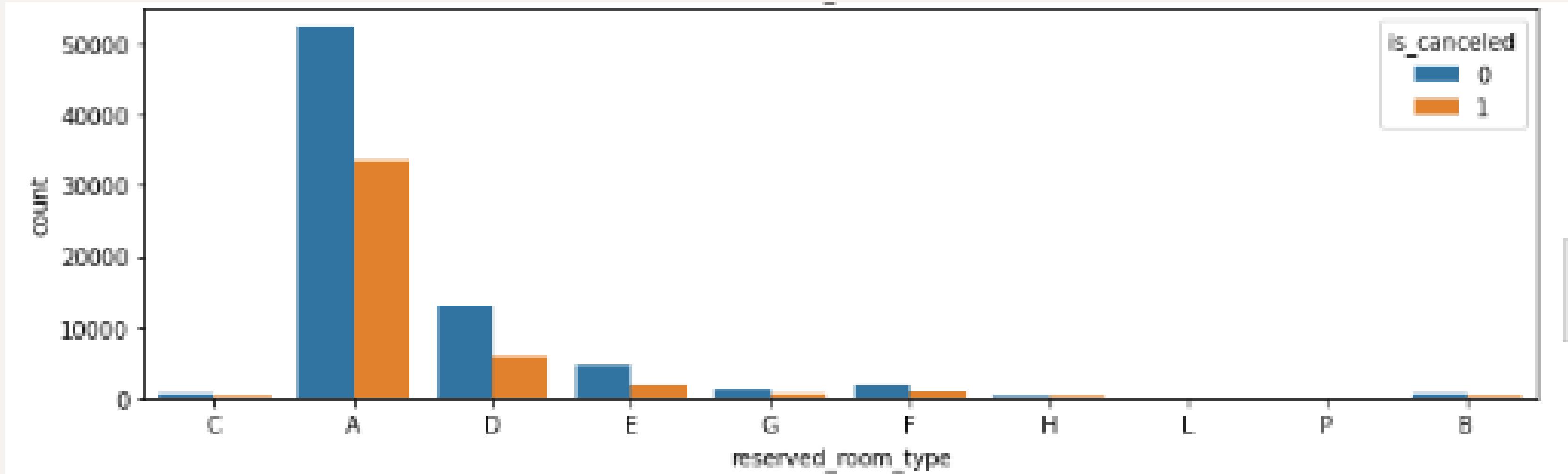
Bulan teramai customer menginap adalah pada bulan Agustus, dan terjadi canceled terbanyak juga pada bulan Agustus.

Market Segmen



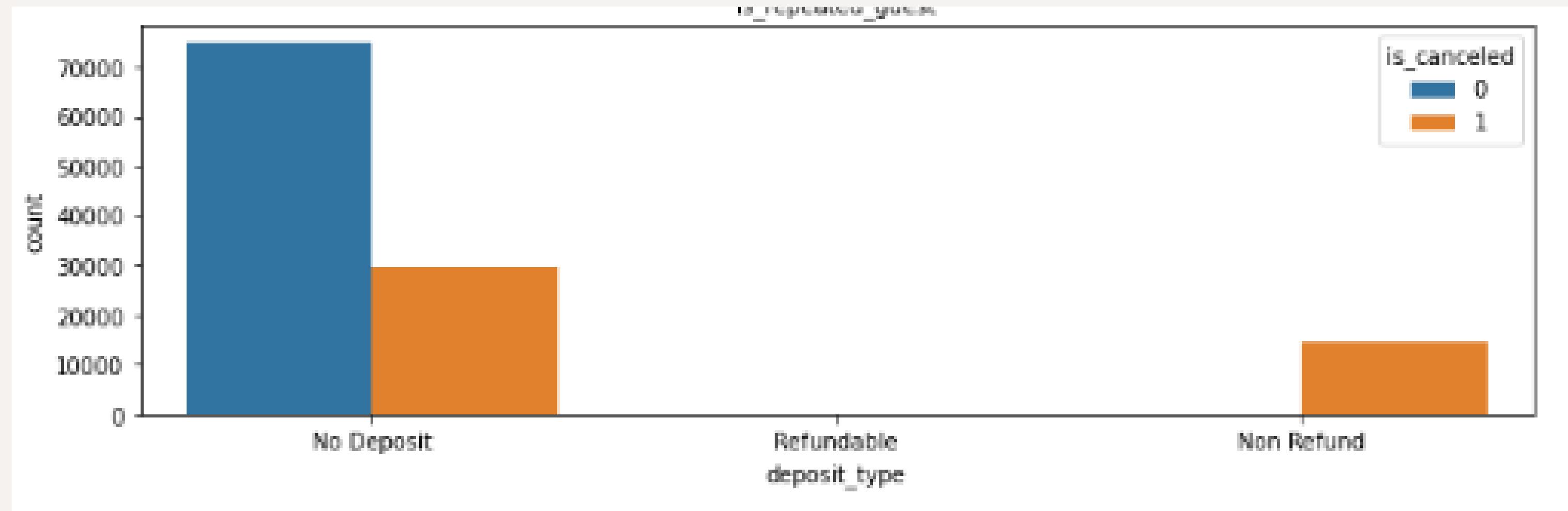
Mayoritas Market segmen ialah Online TA dan pada Complementary tidak ada customer yang melakukan canceled

Reserved room



Mayoritas customer memesan tipe kamar A namun lebih dari setengah customer yang memilih tipe kamar A melakukan canceled.

Deposit type



customer yang memilih non refund melakukan canceled

Data Pre-Processing

1. Check Missing value

- chidren : 4 data
- Agent : 16.340 data
- Company : 112.593 data
- Country : 488 data

2. Mengisi missing value

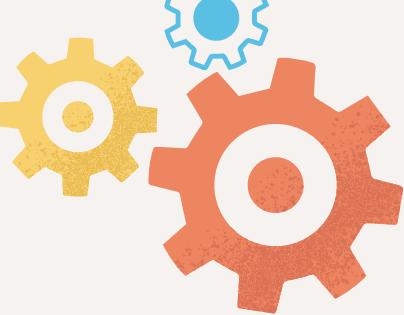
- Kolom children memiliki nilai 0 yang berarti 0 anak-anak hadir dalam kelompok pelanggan yang melakukan transaksi tersebut. Nilai 'nan' adalah nilai yang hilang karena kesalahan pencatatan data. Jadi, kita akan mengganti nilai 'nan' dengan nilai mean dari anak-anak.
- Kolom agent dan company adalah pemesanan yang dilakukan oleh agent atau perusahaan tertentu, mungkin ada beberapa kasus ketika pelanggan tidak memesan hotel melalui agent atau melalui perusahaan mana pun. Jadi dalam hal ini, nilainya bisa nol, sehingga untuk missing value akan diganti dengan nilai nol (0).
- Kolom country adalah kolom yang mewakili negara asal pelanggan. Karena, kolom ini memiliki tipe data string. Jadi, missing value diganti dengan 'others'.



Data Pre-Processing

3. Check Data Unik

hotel	2
is_canceled	2
lead_time	479
arrival_date_year	3
arrival_date_month	12
arrival_date_week_number	53
arrival_date_day_of_month	31
stays_in_weekend_nights	17
stays_in_week_nights	35
adults	14
children	5
babies	5
meal	5
country	177
market_segment	8
distribution_channel	5
is_repeated_guest	2
previous_cancellations	15
previous_bookings_not_canceled	73
reserved_room_type	10
assigned_room_type	12
booking_changes	21
deposit_type	3
agent	333
company	352
days_in_waiting_list	128
customer_type	4
adr	8879
required_car_parking_spaces	5
total_of_special_requests	6
reservation_status	3
reservation_status_date	926
	dtype: int64



Data Pre-Processing

before

```
hotel          0  
is_canceled   0  
lead_time      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  
customer_type   0  
adr            0  
required_car_parking_spaces 0  
total_of_special_requests 0  
reservation_status 0  
reservation_status_date 0  
dtype: int64
```

after

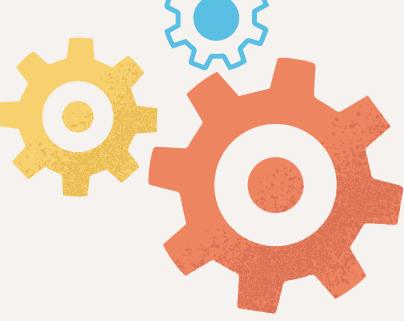
```
hotel          0  
is_canceled   0  
lead_time      0  
arrival_date_month  0  
arrival_date_week_number 0  
arrival_date_day_of_month 0  
stays_in_weekend_nights 0  
stays_in_week_nights    0  
adults         0  
children       0  
babies         0  
meal           0  
country        0  
market_segment  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          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
```



Data Pre-Processing

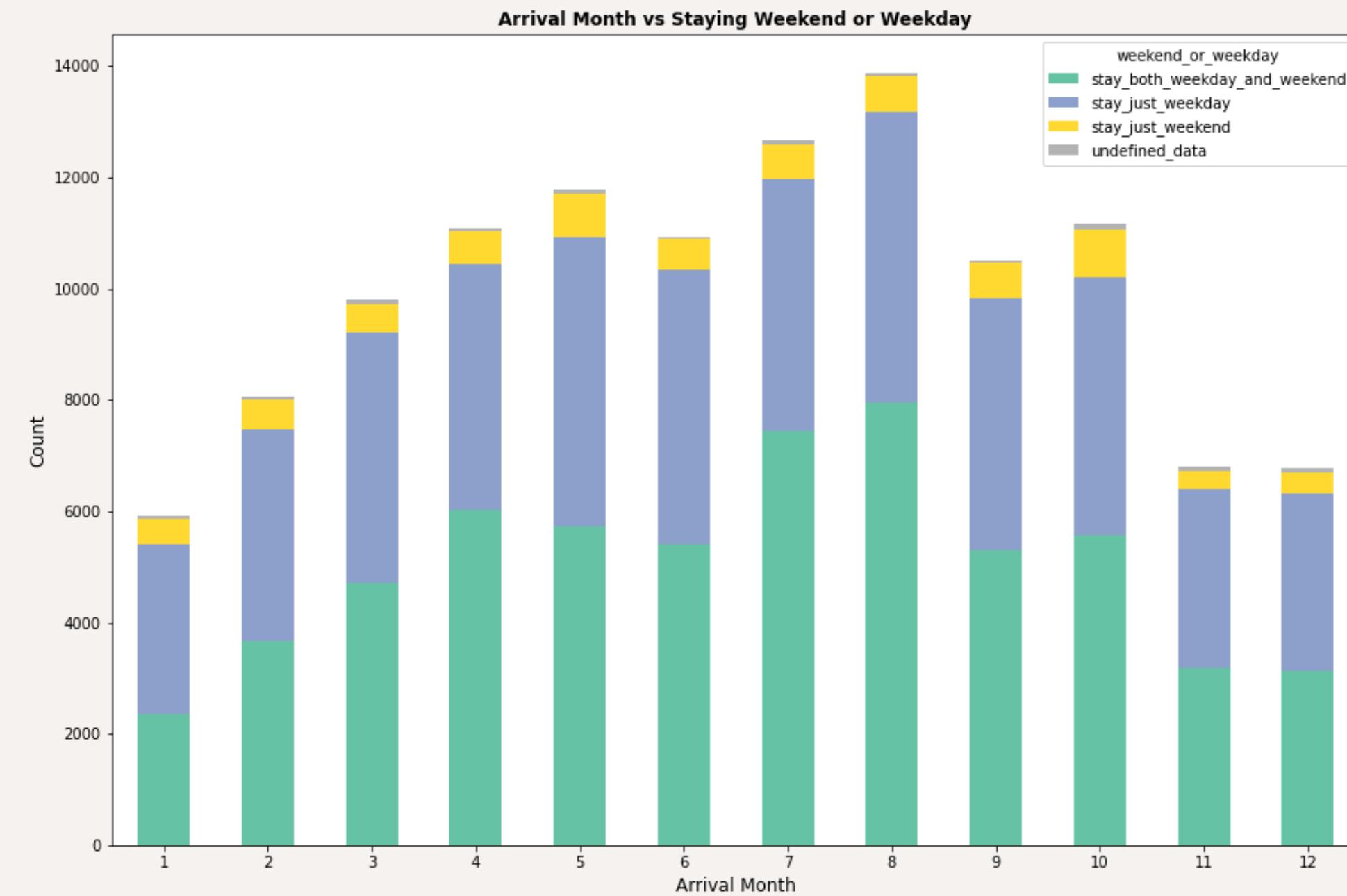
4. Feature engineering

- Weekend_or_Weekdays : untuk memisahkan customer yang menginap saat weekend, menginap saat weekdays atau gabungan keduanya
- All_children : gabungan dari babies dan children



Data Pre-Processing

hubungan fitur 'weekend_or_weekday' dengan 'arrival_date_month'.

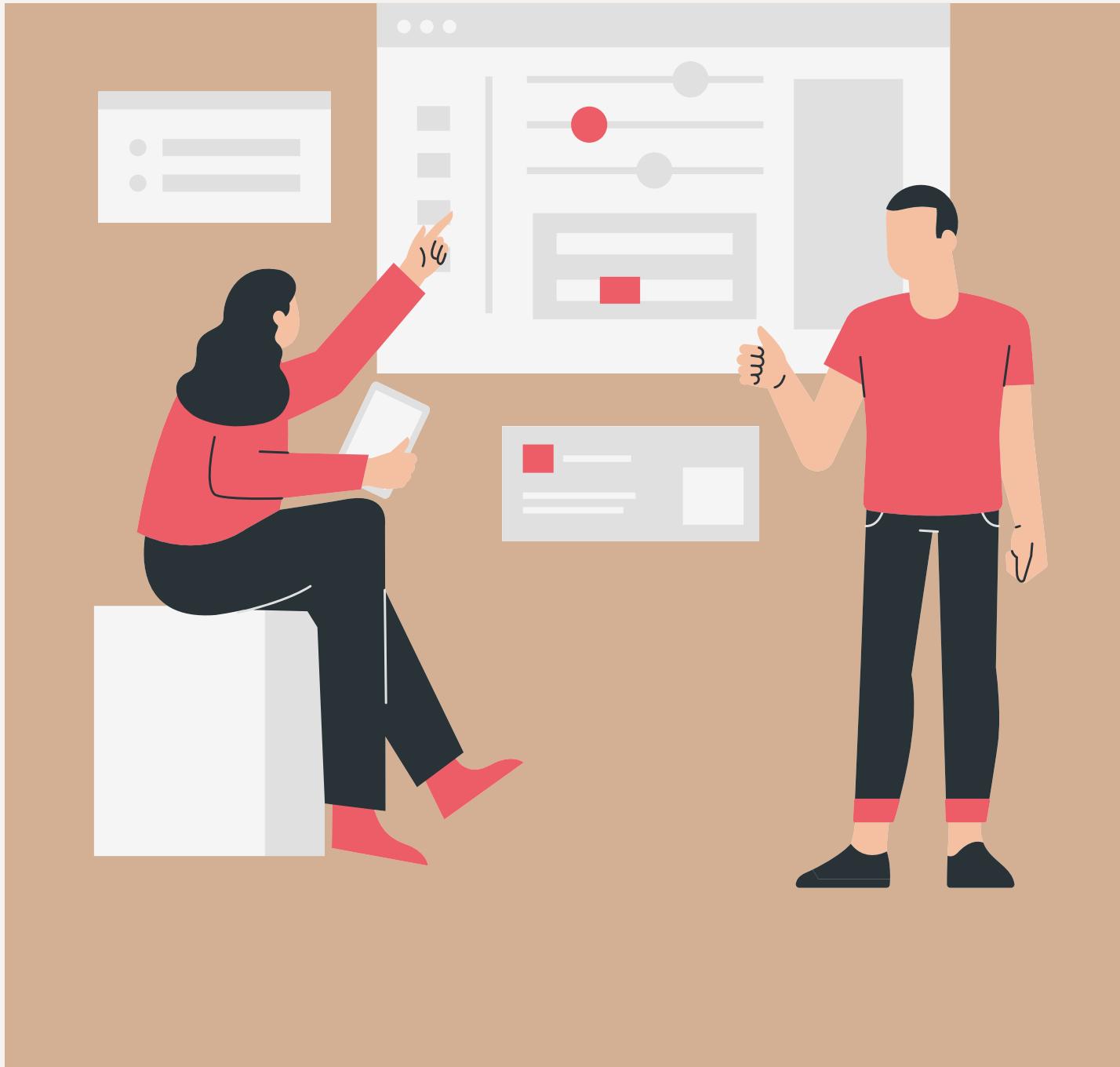


Grafik batang di atas menunjukkan bahwa sebagian besar pemesanan dilakukan untuk menginap hanya untuk hari kerja atau hari kerja dan akhir pekan. Di sisi lain, jumlah stay yang hanya kategori weekend cukup rendah dibandingkan dengan kategori lain.

Modelling

Menggunakan Decision Tree, Random Forest, KNN, dan Logistic Regression.

Sebelum membangun model, data akan dibagi untuk melatih dan menguji rasio masing-masing 80% data training dan 20% data testing.



Decision Tree

```
y_pred_tree = tree.predict(x_test)  
print(classification_report(y_test, y_pred_tree))
```

	precision	recall	f1-score	support
0	0.96	0.96	0.96	15051
1	0.93	0.93	0.93	8827
accuracy			0.95	23878
macro avg	0.94	0.94	0.94	23878
weighted avg	0.95	0.95	0.95	23878

Random Forest

```
y_pred_rf = rf.predict(x_test)
print(classification_report(y_test, y_pred_rf))
```

	precision	recall	f1-score	support
0	0.95	0.99	0.97	15051
1	0.98	0.90	0.94	8827
accuracy			0.96	23878
macro avg	0.96	0.95	0.95	23878
weighted avg	0.96	0.96	0.96	23878

KNN

```
y_pred_knn = knn.predict(x_test)
print(classification_report(y_test, y_pred_knn))
```

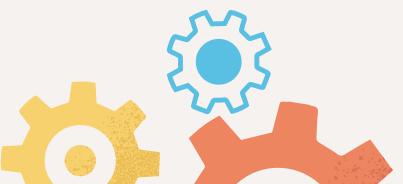
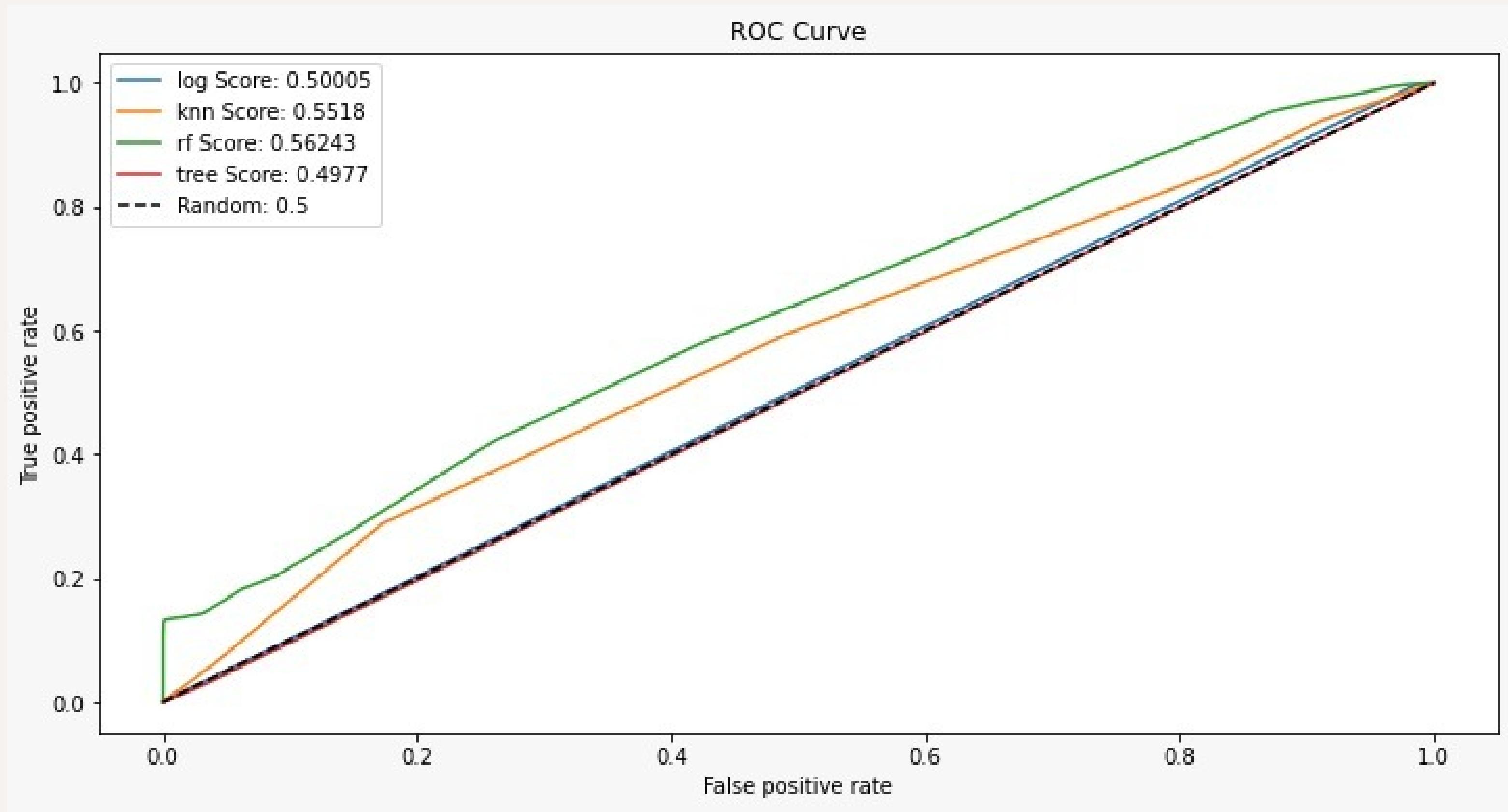
	precision	recall	f1-score	support
0	0.89	0.96	0.92	15051
1	0.92	0.79	0.85	8827
accuracy			0.90	23878
macro avg	0.90	0.87	0.88	23878
weighted avg	0.90	0.90	0.89	23878

Logistic Regression

```
y_pred_log = log.predict(x_test)
print(classification_report(y_test, y_pred_log))
```

	precision	recall	f1-score	support
0	0.79	0.94	0.86	15051
1	0.84	0.59	0.69	8827
accuracy			0.81	23878
macro avg	0.82	0.76	0.78	23878
weighted avg	0.81	0.81	0.80	23878

ROC CURVE



resampling to handle imbalanced dataset

before SMOTE

```
0    60115  
1    35397  
Name: is_canceled, dtype: int64
```

after SMOTE

```
1    60115  
0    60115  
Name: is_canceled, dtype: int64
```

result accuracy SMOTE

		precision	recall	f1-score	support
	0	0.95	0.99	0.97	15051
	1	0.98	0.91	0.94	8827
	accuracy			0.96	23878
	macro avg	0.96	0.95	0.96	23878
	weighted avg	0.96	0.96	0.96	23878

Kesimpulan

- 1) Model yang terbaik berdasarkan akurasi tertinggi adalah Random forest dengan tingkat akurasi sebesar 96%.
- 2) Variabel yang memengaruhi customer yang melakukan cancel adalah variabel 'hotel', 'lead_time', 'arrival_date_week_number', 'arrival_date_day_of_month', 'stays_in_weekend_nights', 'stays_in_week_nights', 'adults', 'meal', 'country', 'market_segment', 'distribution_channel', 'is_repeated_guest', 'previous_cancellations', 'previous_bookings_not_canceled', 'reserved_room_type', 'deposit_type' 'agent', 'company', 'customer_type', 'adr', 'required_car_parking_spaces', 'total_of_special_requests', 'weekend_or_weekday', 'all_children', 'year', 'month', 'day'.

Thankyou