DATA SCIENCE

HOTEL BOOKING CANCELATIONS MODELING

untuk mengidentifikasi faktor-faktor yang memengaruhi pembatalan pemesanan hotel (hotel booking cancellation)

M AGIL MUBAROK	2012210012
AFICA MARSHA NURCAHAYA	2012310002
FERINDRIAKARINA LINTANG HELSADAMARA	2012310007
EIS AGHISNI	2012310701

NEXT SLIDE

MENGAPA FOKUS PADA HOTEL BOOKING CANCELLATIONS?

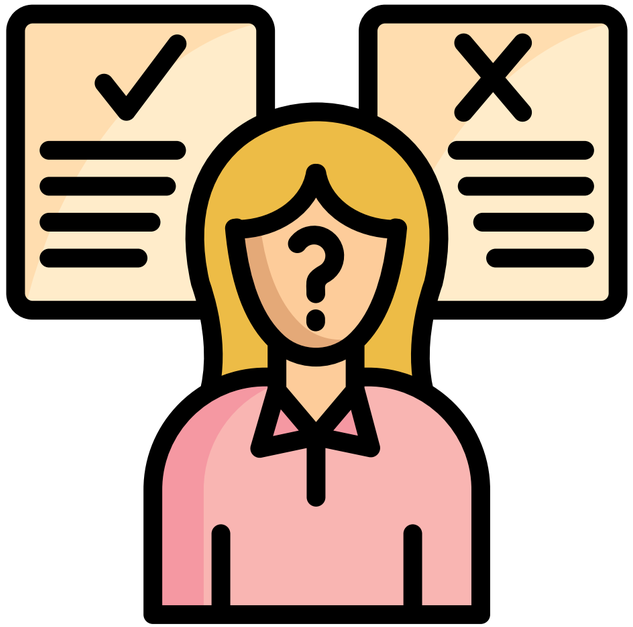
- **Dampak langsung terhadap pendapatan hotel**
  Pembatalan mengurangi tingkat hunian dan menyebabkan kerugian finansial.
- **Peluang untuk prediksi dan pencegahan**
  Dengan data sains, pembatalan bisa diprediksi sehingga hotel dapat mengambil langkah proaktif.
- **Menunjang pengambilan keputusan manajerial**
  Hasil analisis dapat digunakan untuk menyusun strategi promosi, kebijakan refund, dan sistem pemesanan yang lebih efektif.

SUMBER DATA PENELITIAN

	hotel	is_canceled	lead_time	market_segment	is_repeated_guest	previous_cancellations	reserved_room_type	booking_changes	deposit_type	adr	reservation_status_date	name	phone-number
0	Resort Hotel	0	342	Direct	0	0	C	3	No Deposit	0.0	7/1/2015	Ernest Barnes	669-792-1661
1	Resort Hotel	0	737	Direct	0	0	C	4	No Deposit	0.0	7/1/2015	Andrea Baker	858-637-6955
2	Resort Hotel	0	7	Direct	0	0	A	0	No Deposit	75.0	7/2/2015	Rebecca Parker	652-885-2745
3	Resort Hotel	0	13	Corporate	0	0	A	0	No Deposit	75.0	7/2/2015	Laura Murray	364-656-8427
4	Resort Hotel	0	14	Online TA	0	0	A	0	No Deposit	98.0	7/3/2015	Linda Hines	713-226-5883
5	Resort Hotel	0	14	Online TA	0	0	A	0	No Deposit	98.0	7/3/2015	Jasmine Fletcher	190-271-6743
6	Resort Hotel	0	0	Direct	0	0	C	0	No Deposit	107.0	7/3/2015	Dylan Rangel	420-332-5209
7	Resort Hotel	0	9	Direct	0	0	C	0	No Deposit	103.0	7/3/2015	William Velez	286-669-4333

Pengumpulan data yang digunakan dalam modeling hotel booking cancelations, diambil dari situs Kaggle. Data terdiri dari beberapa variabel: hotel, is_canceled, lead_time, market_segment, is_repeated_guest, previous_cancellations, reserved_room_type, booking_changes, deposit_type, adr, reservation_status_date, name, phone-number.

Jumlah data: 119.390 Data



TUJUAN ANALISIS

01 Mengetahui distribusi data pembatalan pemesanan hotel

02 Menganalisis hubungan variabel-variabel lain terhadap pembatalan

03 Mengidentifikasi ketidakseimbangan data antara pemesanan yang dibatalkan dan tidak dibatalkan.

FITUR KATEGORIKAL > NUMERIK

Label Encoder

market_segment > Kategori Kehadiran

```
marketsegment_map = {
    'Direct': 0, 'Corporate': 1, 'Online TA': 2, 'Offline TA/TO': 3,
    'Complementary': 4, 'Groups': 5, 'Undefined': 6, 'Aviation': 7
```

reserved_room_type > Tipe Kamar Dipesan

```
reserved_room_map = {
    'A': 0, 'B': 1, 'C': 2, 'D': 3,
    'E': 4, 'F': 5, 'G': 6, 'H': 7,
    'L': 8
```

deposit_type > Tipe Deposit

```
deposit_type_map = {
    'No Deposit': 0,
    'Refundable': 1,
    'Non Refund': 2
```

DATA NULL HANDLING

```
, coercing errors, then fill NaNs with the mode
pesan'] = pd.to_numeric(dfnew['Tipe Kamar Dipesan'], errors='coerce')
pesan'] = dfnew['Tipe Kamar Dipesan'].fillna(dfnew['Tipe Kamar Dipesan'].mode()[0])
```

```
[ ] dfnew.isnull().sum()
```

	0
hotel	0
is_canceled	0
lead_time	0
market_segment	0
is_repeated_guest	0
previous_cancellations	0
reserved_room_type	0
booking_changes	0
deposit_type	0
adr	0
reservation_status_date	0
name	0
phone-number	0
Kategori Kehadiran	0
Tipe Kamar Dipesan	12
Tipe Deposit	0

dtype: int64

```
[ ] dfnew.isnull().sum()
```

	0
hotel	0
is_canceled	0
lead_time	0
market_segment	0
is_repeated_guest	0
previous_cancellations	0
reserved_room_type	0
booking_changes	0
deposit_type	0
adr	0
reservation_status_date	0
name	0
phone-number	0
Kategori Kehadiran	0
Tipe Kamar Dipesan	0
Tipe Deposit	0

dtype: int64

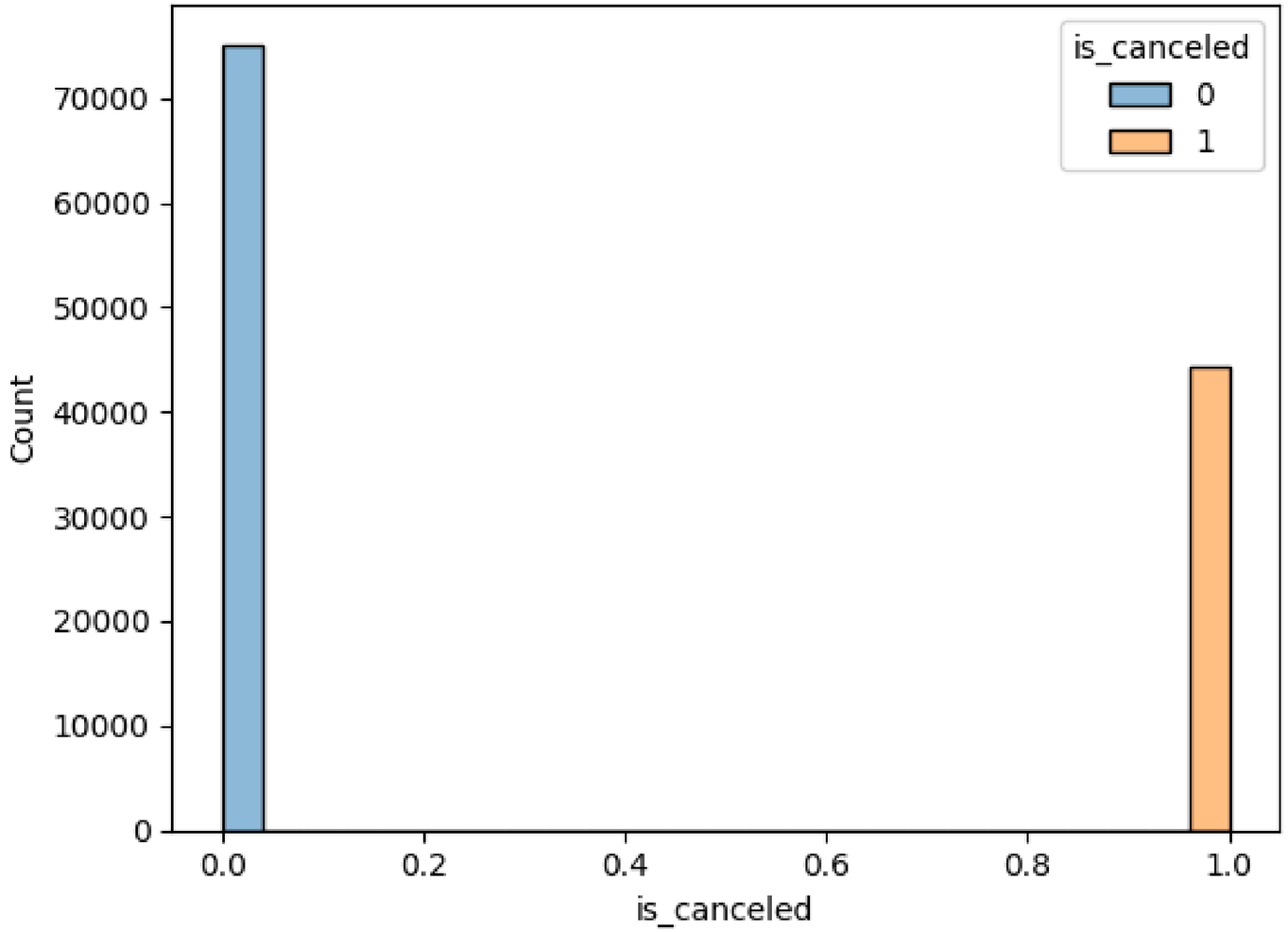
LANGKAH PENGOLAHAN DATA

EDA
Data Preprocessing
Modeling

EDA

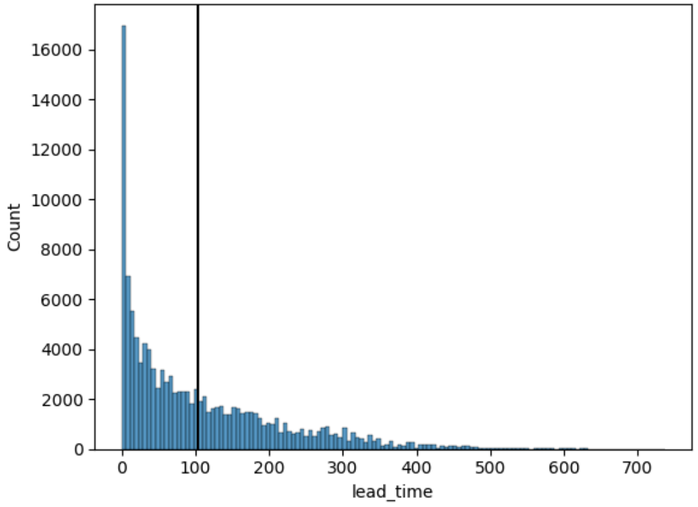
Univariate Data Analysis

Histogram menunjukkan bahwa sebagian besar tamu tidak membatalkan pesanannya, sementara yang membatalkan hanya sedikit. Karena jumlahnya tidak seimbang, data perlu balancing pada tahap preprocessing agar model tidak bias.

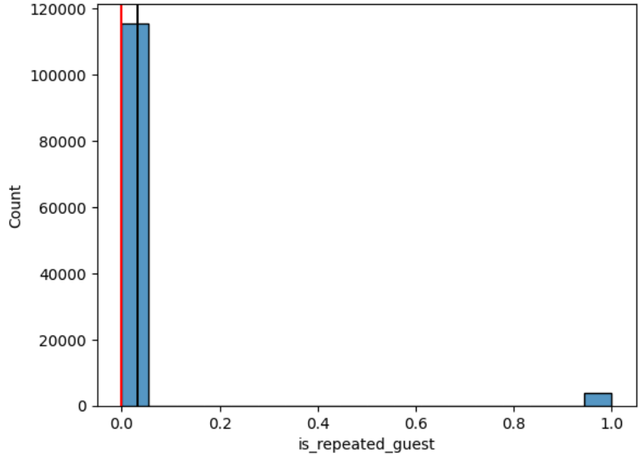


EDA

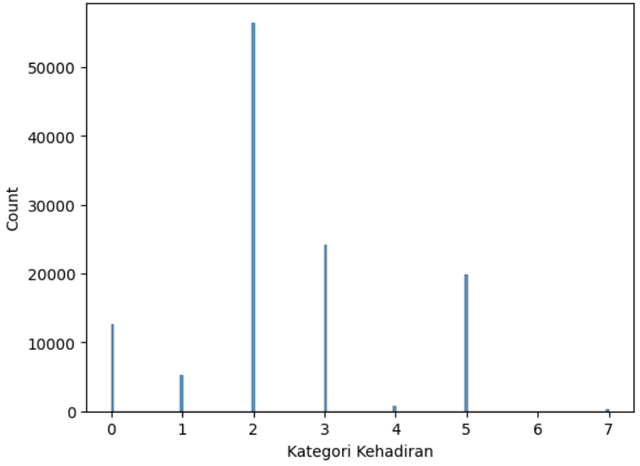
Univariate Data Analysis



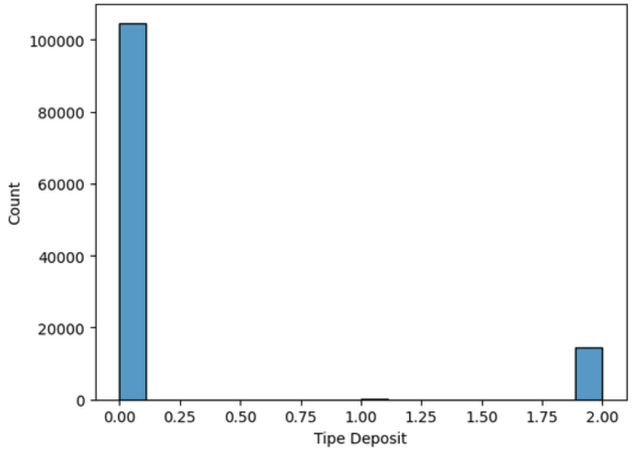
Semakin besar lead time, semakin tinggi kemungkinan pembatalan,



cenderung bukan tamu berulang yang melakukan pembatalan



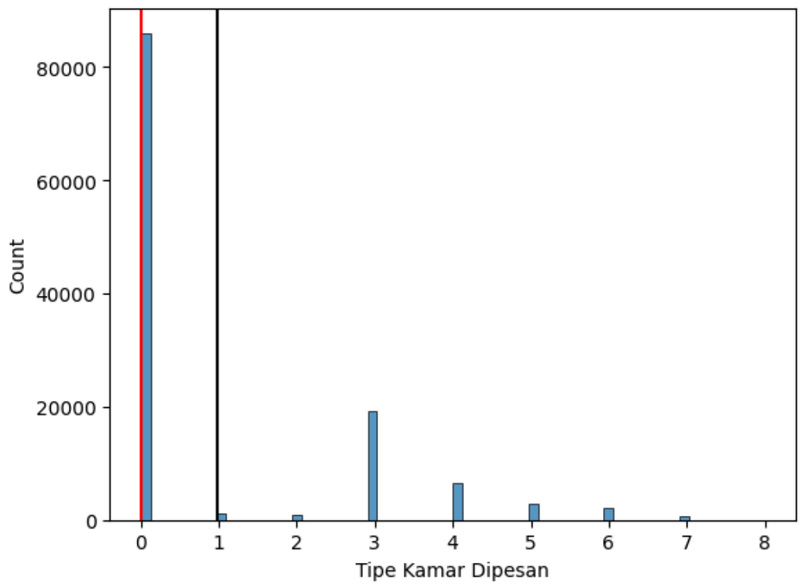
Online TA (Kode 2) adalah kategori terbanyak, artinya mayoritas tamu memesan melalui agen perjalanan online



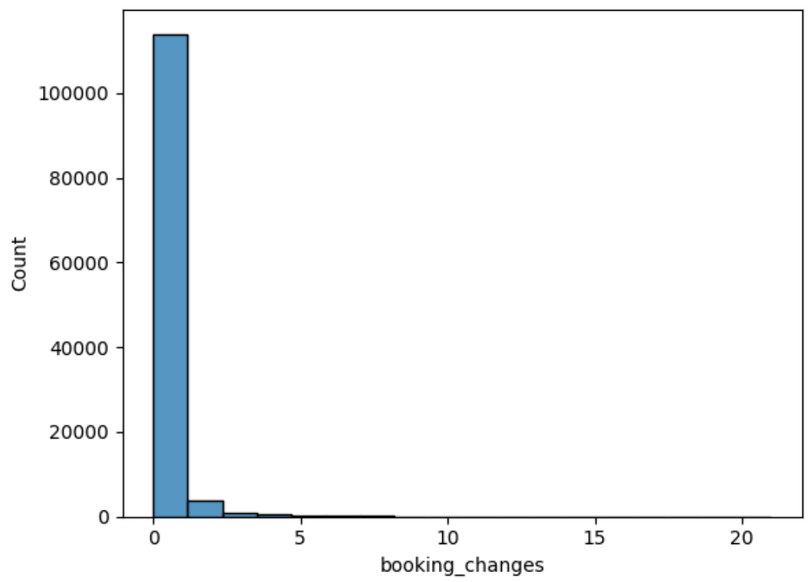
customer cenderung reservasi tanpa deposit, jadi dengan mudah membatalkan karena tidak ada kerugian

EDA

Univariate Data Analysis



Kamar tipe A (Kode 0) adalah kategori terbanyak, artinya banyak customer yang reservasi tipe kamar ini.

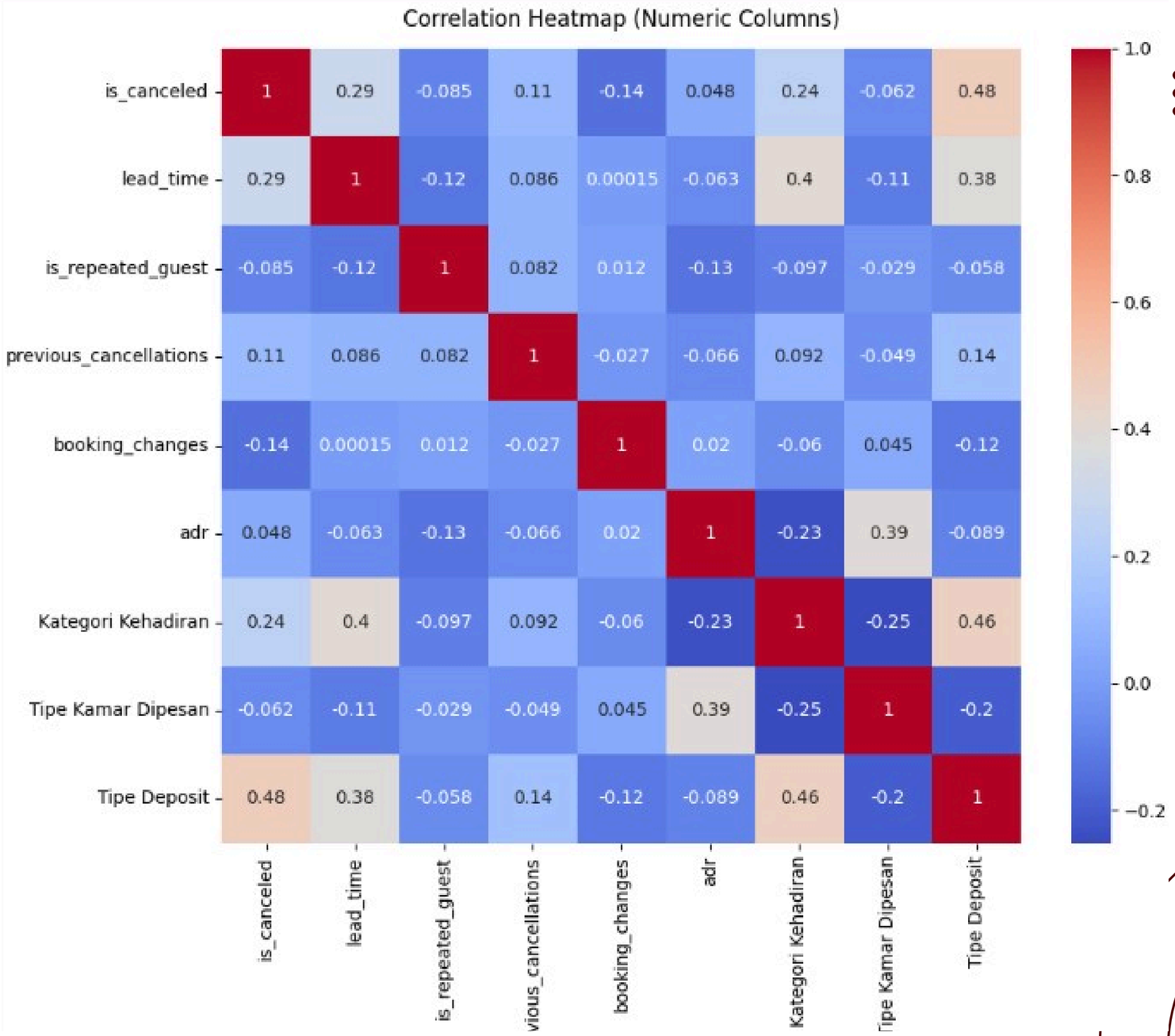


tamu cenderung tidak melakukan banyak perubahan, sehingga tamu pasti hadir

EDA

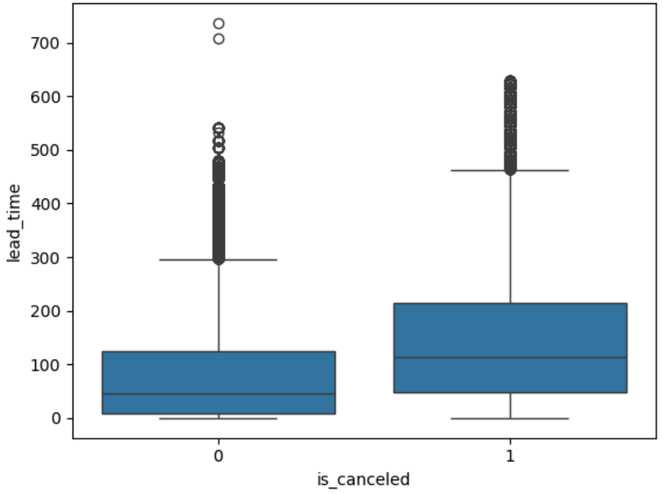
Bivariate Data Analysis

Ada korelasi positif sedang antara is_canceled dengan lead_time. Artinya, semakin jauh waktu pemesanan dari tanggal menginap, semakin besar kemungkinan pemesanan akan dibatalkan.

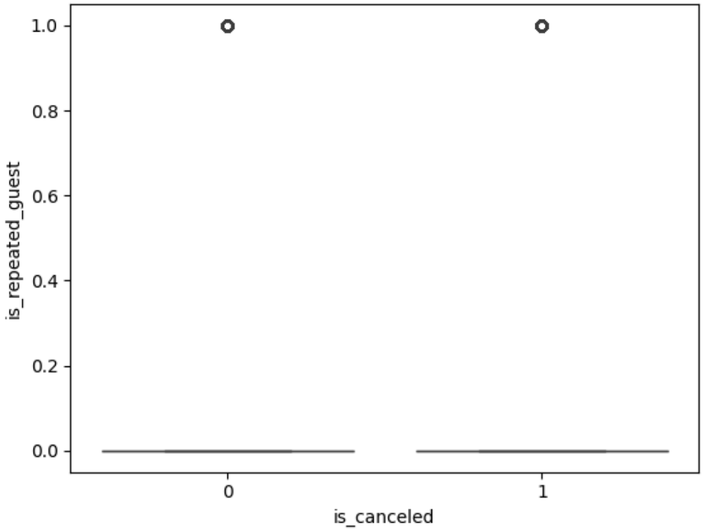


EDA

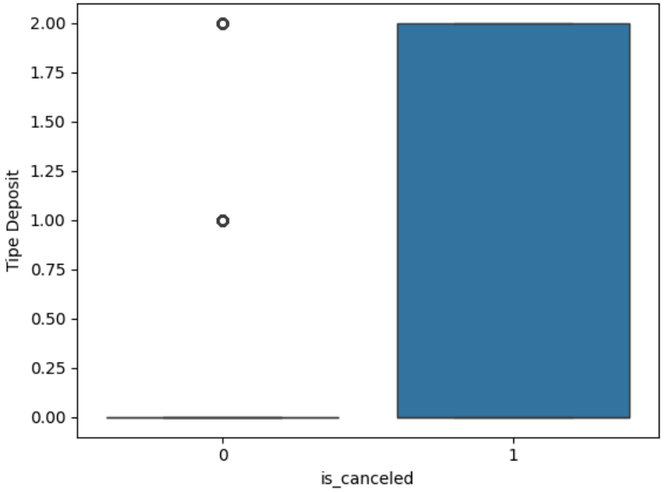
Bivariate Data Analysis



Semakin besar lead time, semakin tinggi kemungkinan pembatalan,



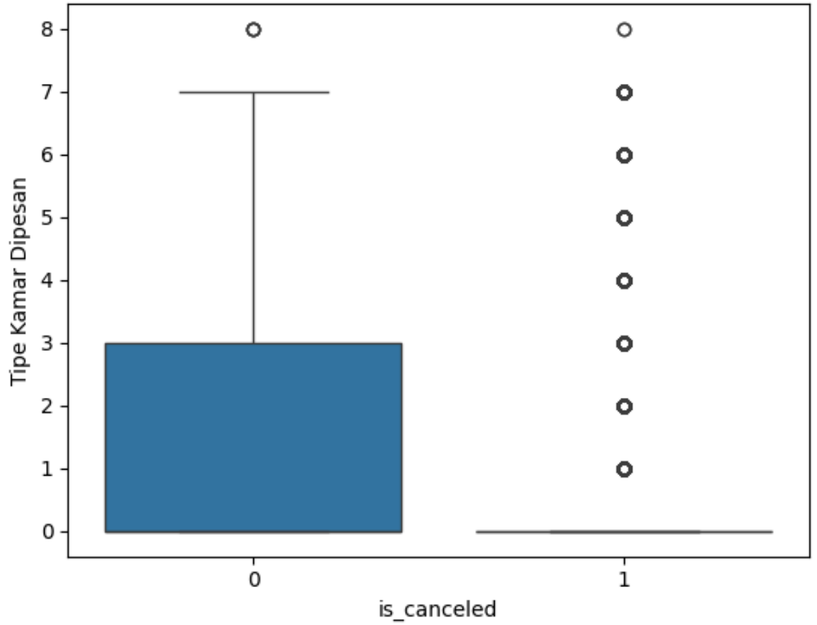
Tamu berulang lebih loyal dan memiliki kemungkinan kecil untuk membatalkan reservasi.



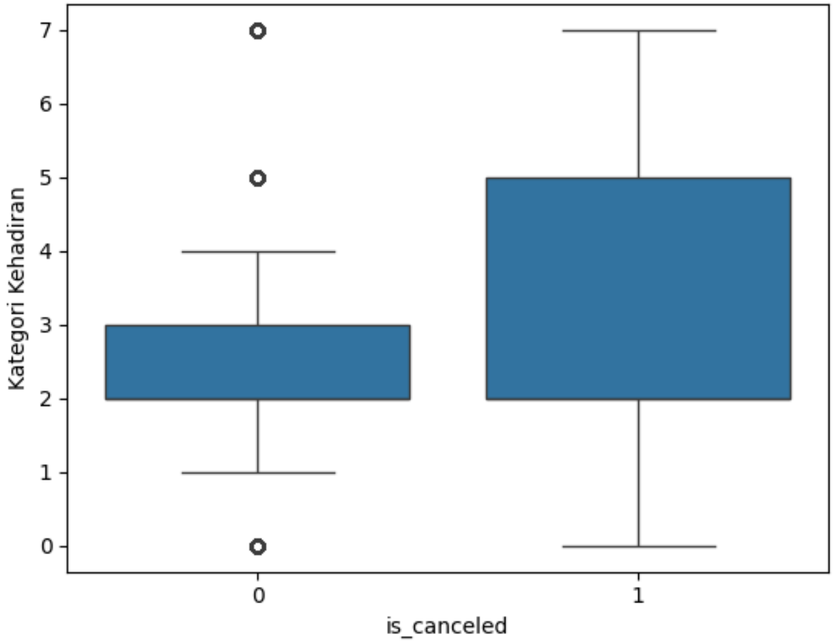
Pelanggan tanpa komitmen keuangan lebih cenderung membatalkan pemesanan

EDA

Bivariate Data Analysis



Tipe kamar lebih mahal atau spesifik cenderung jarang dibatalkan.

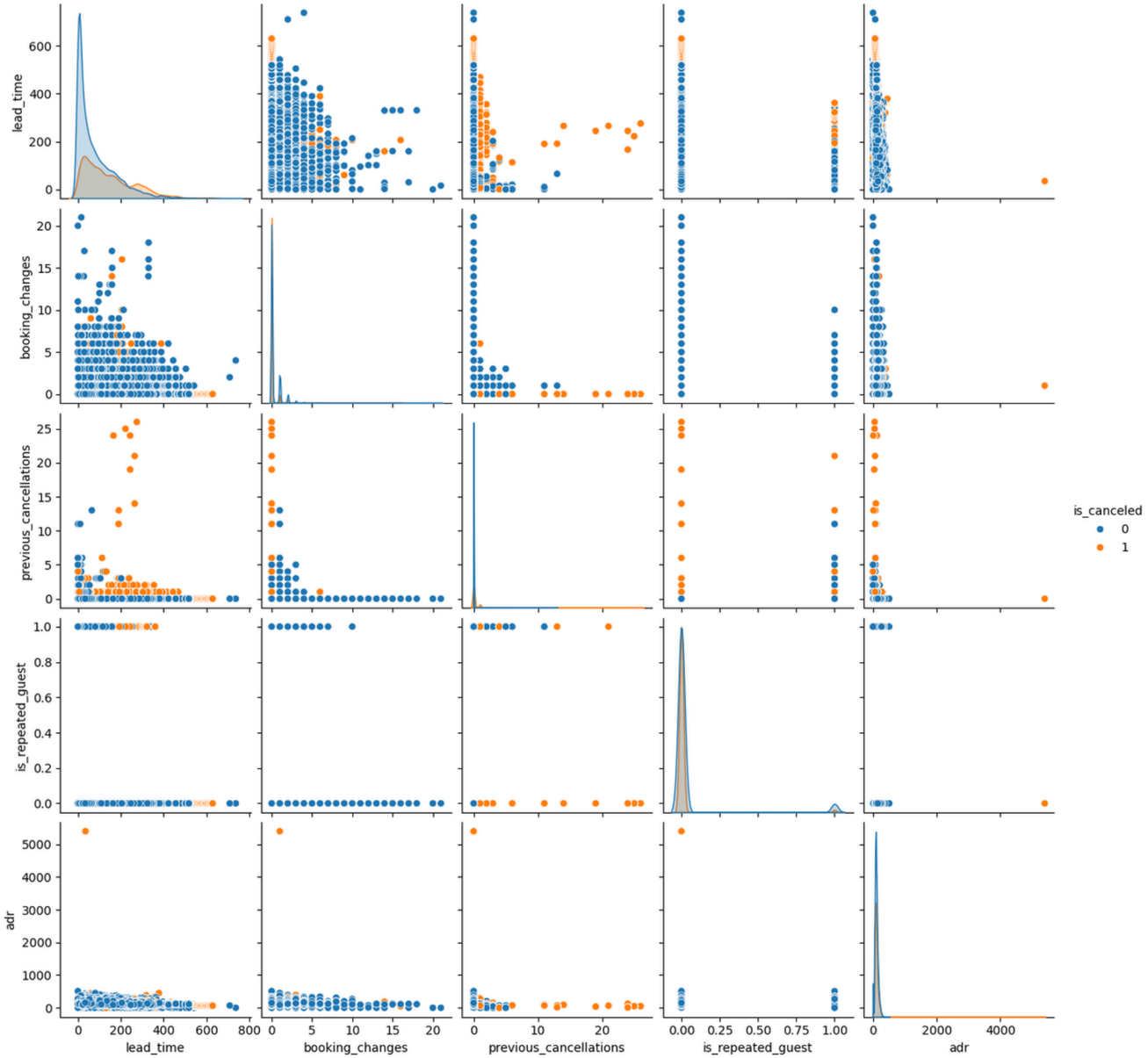


Sebagian besar pembatalan dilakukan oleh pelanggan dengan kunjungan rendah, meskipun ada juga pelanggan dengan kunjungan tinggi yang membatalkan.

EDA

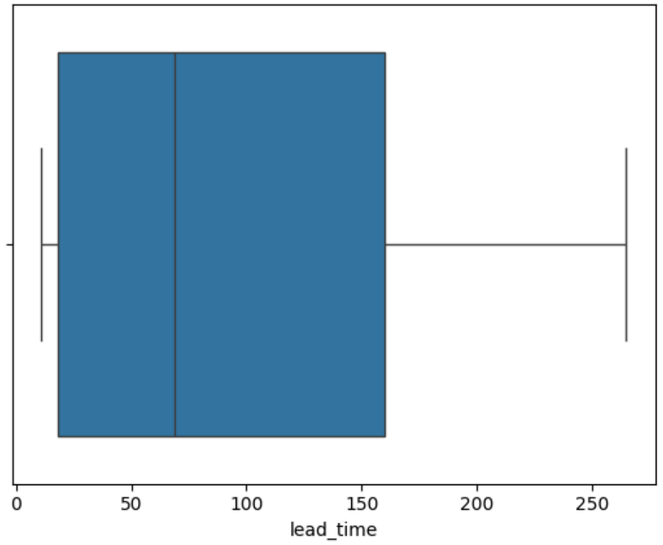
Multivariate Data Analysis

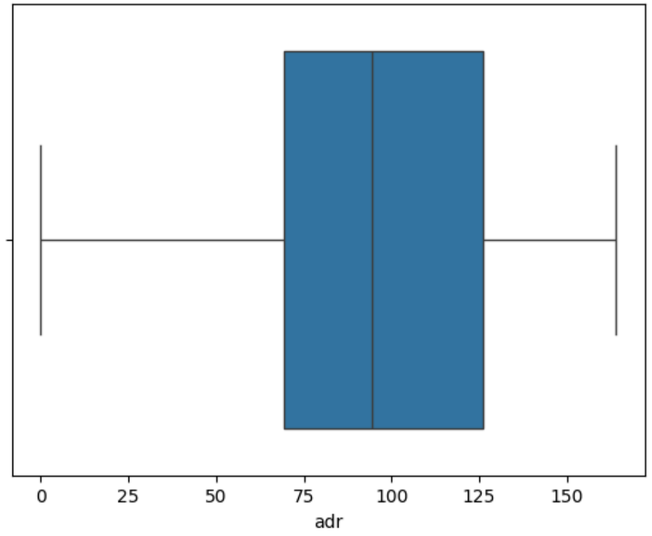
Tamu yang memesan jauh-jauh hari (lead_time tinggi) dan yang pernah membatalkan sebelumnya lebih sering membatalkan lagi. Pelanggan baru lebih banyak membatalkan dibanding pelanggan lama. Pembatalan juga cenderung terjadi pada harga tertentu yang tinggi (adr). Sementara itu, variabel booking_changes tidak menunjukkan pengaruh yang jelas. Variabel lead_time, previous_cancellations, dan adr terlihat paling berpengaruh terhadap pembatalan.

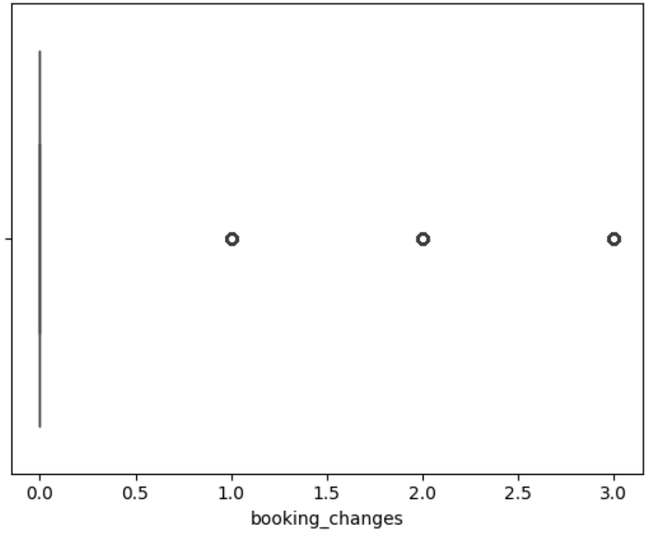


DATA PREPROCESSING

OUTLIER HANDLING







Lead_Time

Banyak pelanggan yang memesan jauh hari (hingga > 150 hari), tapi juga ada yang memesan mendekati tanggal menginap (< 20 hari).

adr

Range besar menunjukkan variasi tarif tergantung tipe hotel atau kondisi lainnya.

Booking_Change

Mayoritas pelanggan tidak melakukan perubahan pemesanan (bernilai 0). Terdapat outlier di nilai 1, 2, dan 3, Artinya sebagian kecil pelanggan melakukan revisi terhadap booking awal mereka.

DATA PREPROCESSING

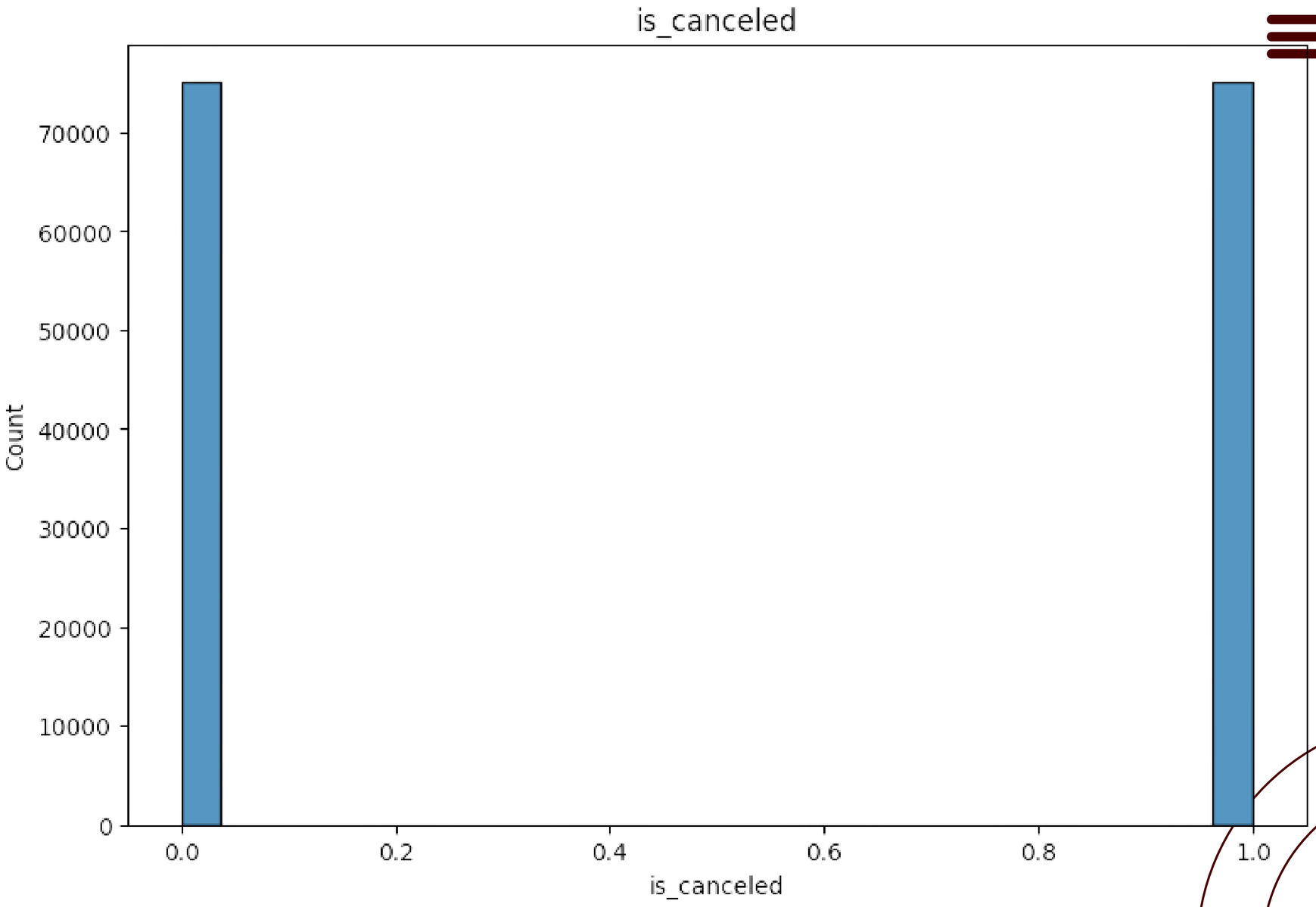
NORMALIZATION

MinMax Scaler

```
Data yang sudah di-Min-Max scaler:
[[0. 0. 1. ... 0. 0. 1.]
 [0. 0. 1. ... 0. 0. 1.]
 [0. 0. 0. ... 0. 0. 1.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 1.]]
```

Proses mengubah data ke dalam rentang tertentu 0 sampai 1 untuk memastikan semua fitur memiliki skala yang seragam.

IMBALANCE DATA HANDLING



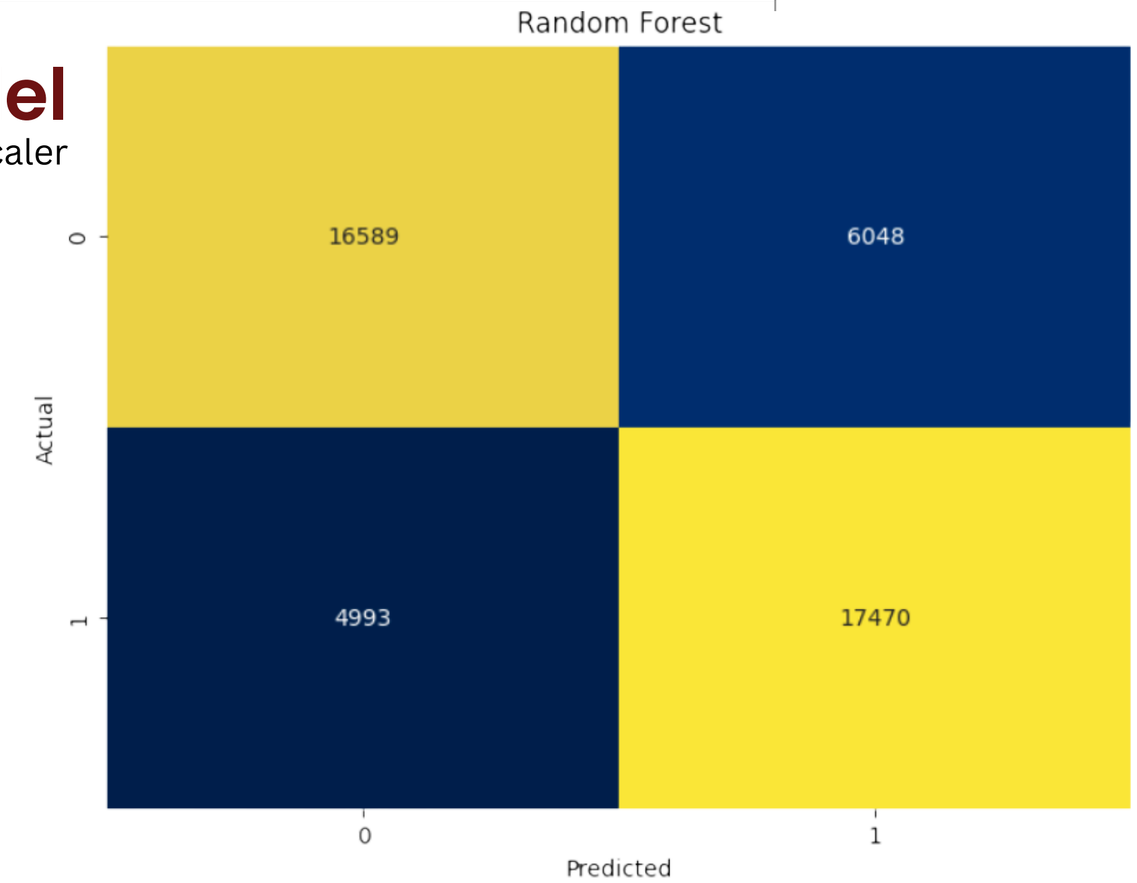
MODELLING

Model	Performance (%)			
	Accuracy	Precision	Recall	F1-Score
XGBoost	0.73	0.74	0.71	0.72
Logistic Regression	0.68	0.80	0.49	0.60
Random Forest	0.75	0.74	0.77	0.75

Best Model

dengan pendekatan MinMaxScaler

	precision	recall	f1-score	support
0	0.73	0.77	0.75	21582
1	0.78	0.74	0.76	23518
accuracy			0.76	45100
macro avg	0.76	0.76	0.76	45100
weighted avg	0.76	0.76	0.76	45100



KESIMPULAN

Tipe deposit sangat berpengaruh: pelanggan tanpa deposit (No Deposit) mungkin lebih sering membatalkan.

Fitur booking_changes dan previous_cancellations bisa menjadi indikator perilaku pelanggan yang tidak konsisten.

Pelanggan dengan lead_time yang tinggi berpotensi membatalkan lebih besar.

Fitur adr (harga kamar) juga dapat berperan, apalagi jika harga tinggi.

SEKIAN

TERIMA KASIH

Sekali lagi, terima kasih atas kesempatan ini. Mari kita lanjutkan diskusi ini di lain waktu. Sampai jumpa!

SELESAI