Laporan Mata Kuliah Pembelajaran Mesin

HOTEL BOOKING DEMAND

disusun untuk memenuhi tugas

mata kuliah Pembelajaran Mesin

Oleh :

Kelompok 3

Anggota :

Meutia Aini	(2208107010005)
Akhsania Maisa Rahmah	(2208107010017)
Fadli Ahmad Yazid	(2208107010032)
Muhammad Mahathir	(2208107010056)
Muhammad Aufa Zaikra	(2208107010070)

JURUSAN INFORMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM UNIVERSITAS SYIAH KUALA

DARUSSALAM, BANDA ACEH

2025

A. Data Description

Dataset yang digunakan adalah Hotel Booking Demand yang bersumber dari Kaggle. Dataset ini berisi informasi terkait pemesanan kamar di dua jenis hotel, yaitu City Hotel dan Resort Hotel. Data mencakup pemesanan yang tiba antara 1 Juli 2015 dan 31 Agustus 2017, termasuk pemesanan yang berhasil check in dan pemesanan yang dibatalkan.

Dataset Hotel Booking Demand terdiri dari 119.390 baris data, dengan setiap baris mewakili satu pemesanan hotel. Dataset ini memiliki 32 kolom fitur yang mencakup informasi tentang jenis hotel, status pembatalan, tanggal kedatangan, lama menginap, jumlah tamu, serta berbagai faktor lain yang relevan dalam industri perhotelan.

informasi lainnya yang relevan. Dataset ini tersedia dalam format CSV (Comma-Separated Values), sehingga dapat dengan mudah diolah menggunakan berbagai alat analisis data seperti Python (pandas), R, atau software spreadsheet untuk eksplorasi dan pemodelan lebih lanjut.

Link Dataset: https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand/data

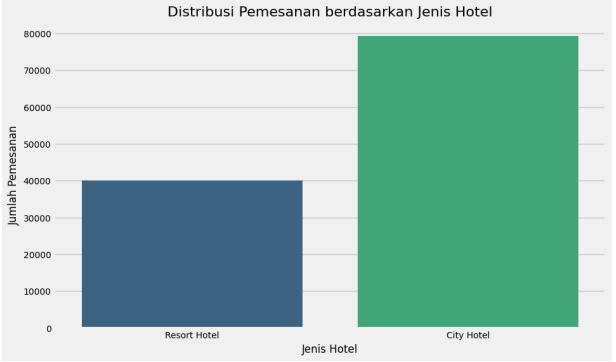
B. Data Loading

Pada bagian ini memuat dataset Hotel Booking Demand akan ke dalam lingkungan pemrograman Python. Proses ini menggunakan library Pandas untuk memuat dan memanipulasi data. Dataset menggunakan fungsi read_csv() dari Pandas. Dataset tersedia dalam format CSV, sehingga mudah dimuat menggunakan Pandas. Contoh kode nya yaitu:

```python
import pandas as pd
import requests
from io import StringIO

url = "https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-02-11/hotels.csv"

response = requests.get(url)
df = pd.read_csv(StringIO(response.text))
df.head()
```

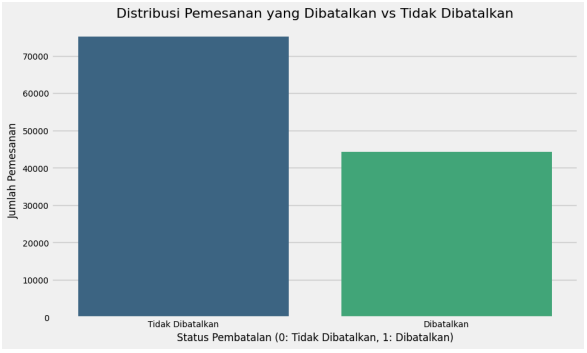
C. Data Understanding

1. Visualisasi Data
    - Distribusi Pemesanan berdasarkan Jenis Hotel



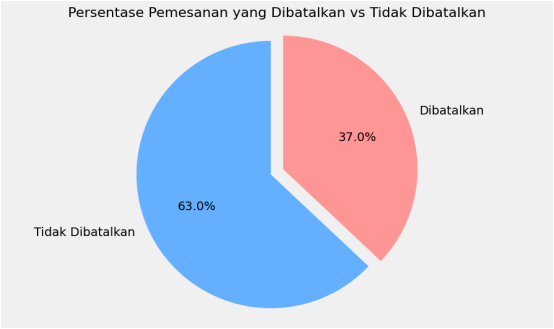
Grafik menunjukkan pemesanan City Hotel lebih tinggi dari Resort Hotel, dengan lebih dari 70.000 vs. sekitar 40.000 pemesanan. Ini menandakan hotel di perkotaan lebih diminati.

- Distribusi Pemesanan yang Dibatalkan vs Tidak Dibatalkan



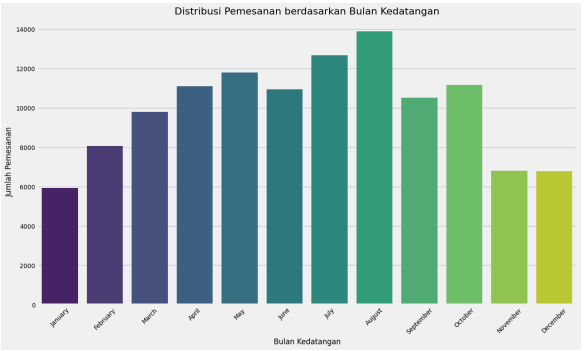
Grafik menunjukkan pemesanan yang tidak dibatalkan lebih banyak (~75.000) dibanding yang dibatalkan (~45.000), meski tingkat pembatalan signifikan, mayoritas pelanggan tetap melanjutkan pesanan.

- Persentase Pemesanan yang Dibatalkan vs Tidak Dibatalkan



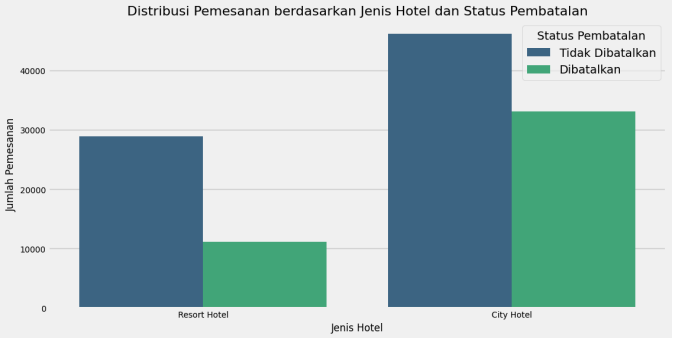
Grafik ini menunjukkan bahwa 37% pemesanan dibatalkan, sementara 63% tetap berlanjut. Tingkat pembatalan yang cukup signifikan ini dapat memengaruhi operasional hotel.

- Distribusi Pemesanan berdasarkan bulan Kedatangan



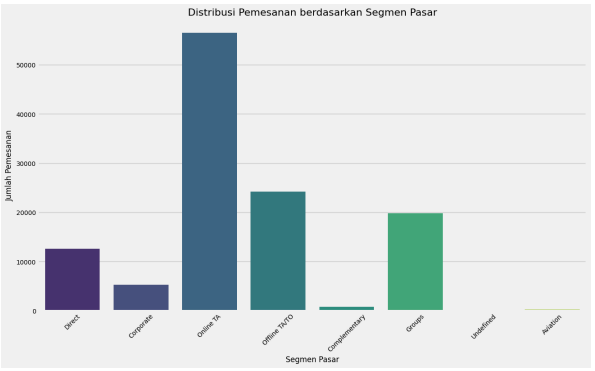
Grafik menunjukkan pemesanan hotel meningkat sejak awal tahun, puncak di Agustus, lalu menurun di akhir tahun, mengindikasikan pola musiman.

- Distribusi Pemesanan berdasarkan Jenis Hotel dan Status Pembatalan



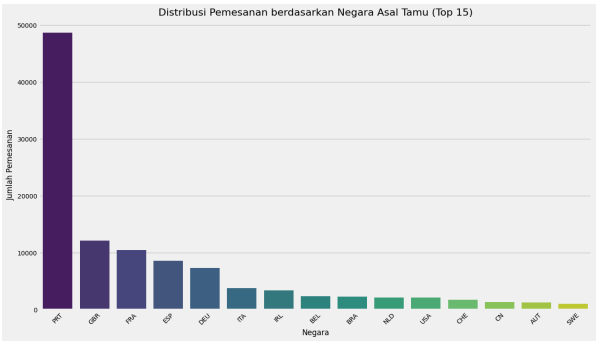
Grafik menunjukkan City Hotel memiliki lebih banyak pemesanan dan tingkat pembatalan lebih tinggi dibanding Resort Hotel, menandakan pemesanannya lebih rentan dibatalkan.

- Distribusi Pemesanan berdasarkan Segmen Pasar



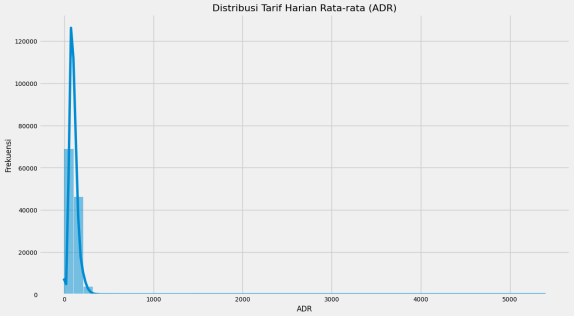
Grafik menunjukkan pemesanan didominasi oleh Online TA, diikuti Offline TA/TO dan Groups. Segmen lainnya jauh lebih kecil, menandakan agen perjalanan online paling banyak digunakan.

- Distribusi Pemesanan berdasarkan Negara Asal Tamu



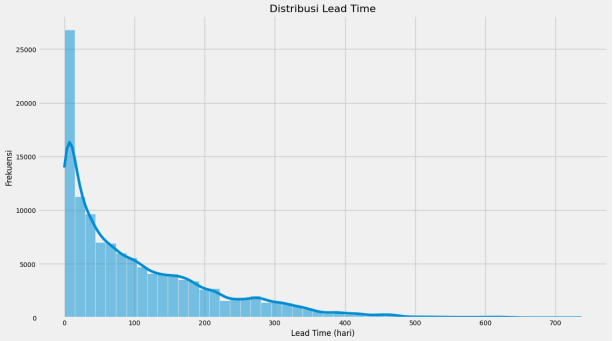
Grafik menunjukkan 15 negara dengan pemesanan tertinggi, dengan satu negara mendominasi. Ini mengindikasikan mayoritas tamu berasal dari negara tersebut, dipengaruhi faktor geografis, kebijakan, atau preferensi wisatawan.

- Distribusi Tarif Harian Rata-rata (ADR)



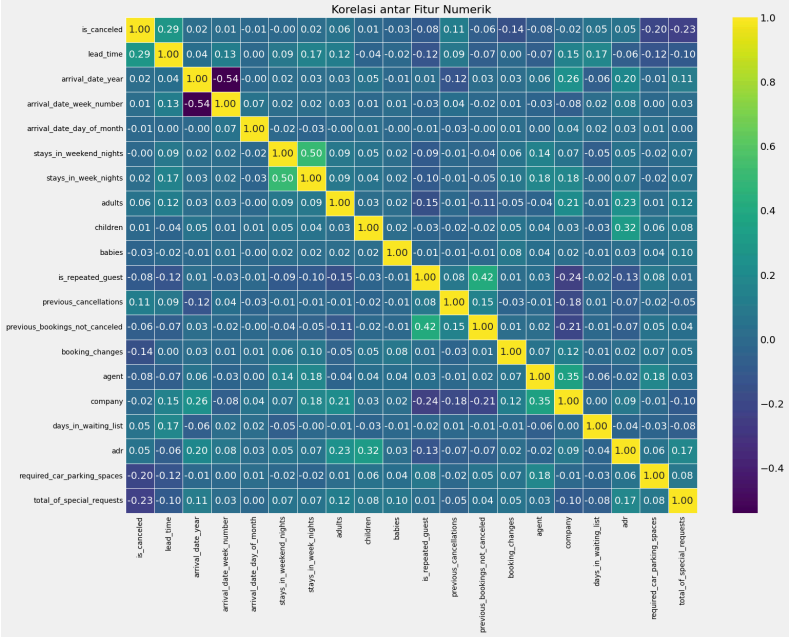
Grafik menunjukkan distribusi ADR skewed ke kanan, dengan mayoritas tarif di bawah 300 dan beberapa outlier tinggi, kemungkinan karena reservasi khusus atau kesalahan data.

- Distribusi Lead Time



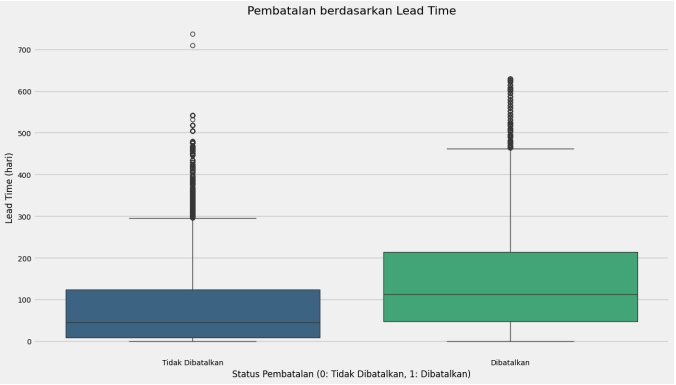
Grafik menunjukkan mayoritas pemesanan dilakukan mendekati tanggal menginap, dengan frekuensi tertinggi pada lead time nol hari, sementara pemesanan jauh sebelumnya lebih jarang.

- Korelasi antar Fitur Numerik



Matriks korelasi menunjukkan lead time berkorelasi positif dengan is_canceled (0.29), indikasi pemesanan lebih awal cenderung dibatalkan. Stays_in_week_nights dan stays_in_weekend_nights berkorelasi (0.50), menunjukkan tamu sering menginap di keduanya. Korelasi lain terlihat pada previous_cancellations & previous_bookings_not_canceled (0.42) serta agent & company (0.35). Secara keseluruhan, korelasi relatif lemah tetapi tetap informatif.

- Analisis Pembatalan berdasarkan Lead Time



Boxplot menunjukkan lead time lebih tinggi untuk reservasi yang dibatalkan, dengan banyak outlier. Ini mengindikasikan pemesanan jauh hari lebih berisiko dibatalkan.

- Lama Menginap berdasarkan Jenis Hotel

Boxplot menunjukkan tamu resort hotel menginap lebih lama dengan variasi lebih besar, sementara city hotel cenderung memiliki durasi menginap lebih pendek dan seragam.

Insight dari Eksplorasi Data:

- **Jenis Hotel:** City Hotel lebih banyak dipesan dibanding Resort Hotel.
- **Pembatalan:** Sekitar 37% pemesanan dibatalkan, berdampak pada pendapatan.
- **Pola Musiman:** Puncak pemesanan terjadi di Juli-Agustus.
- **Segmen Pasar:** "Online TA" mendominasi pemesanan.
- **Negara Asal:** Mayoritas tamu berasal dari Portugal dan negara Eropa lainnya.
- **Tarif (ADR):** Variasi harga besar dengan beberapa outlier tinggi.
- **Lead Time:** Pemesanan jauh-jauh hari lebih berisiko dibatalkan.
- **Lama Menginap:** Mayoritas tamu menginap 1-4 malam, dengan pola berbeda antar hotel.
- **Korelasi:** Lead time berkorelasi positif dengan pembatalan, tamu berulang lebih jarang membatalkan.
- **Variasi Harga:** Resort Hotel lebih fluktuatif dibanding City Hotel.

D. Data Preparation

1. Handling Missing Values
    a. **Menghapus** 'company' → terlalu banyak nilai hilang, kurang berguna.
    b. **Mengisi** 'agent' & 'children' dengan **0** → diasumsikan tanpa agen atau anak.
    c. **Mengisi** 'country' dengan **modus** → tamu terbanyak dianggap mewakili.

    Selain menangani nilai yang hilang, kita juga perlu menangani kasus di mana jumlah tamu (adults + children + babies) adalah 0, yang merupakan anomali dalam data.

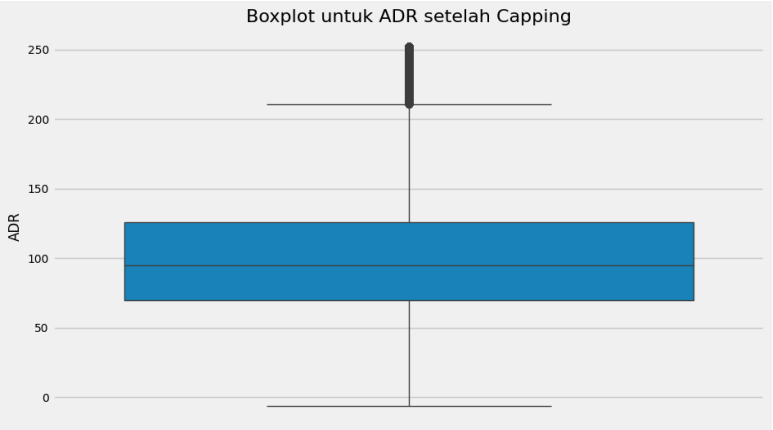
```
# Menghitung jumlah baris di mana jumlah tamu adalah 0
zero_guests = df_prep[(df_prep['adults'] + df_prep['children'] + df_prep['babies']) == 0].shape[0]
print(f"Jumlah baris di mana jumlah tamu adalah 0: {zero_guests}")

# Menghapus baris di mana jumlah tamu adalah 0
df_prep = df_prep[(df_prep['adults'] + df_prep['children'] + df_prep['babies']) > 0]
print(f"Jumlah baris setelah menghapus baris dengan jumlah tamu 0: {df_prep.shape[0]}")
```

Alasan Penghapusan Baris dengan Jumlah Tamu 0:

1. Baris dengan jumlah tamu 0 merupakan anomali karena pemesanan tanpa tamu tidak masuk akal. Menghapusnya adalah langkah yang tepat untuk menghindari distorsi analisis, terutama karena jumlahnya kecil sehingga tidak memengaruhi statistik secara signifikan. Selain itu, mempertahankan data yang tidak valid dapat menyebabkan kesimpulan yang keliru dalam analisis atau model prediktif.

2. Handling Outlier
    - Menangani outlier pada kolom 'adr' dengan metode capping, yaitu membatasi nilai maksimum pada persentil ke-99.
    - Alasan: Metode ini memungkinkan kita untuk menangani nilai ekstrem tanpa menghapus terlalu banyak data, sehingga mempertahankan informasi yang berharga.

    Kolom 'adr' memiliki beberapa outlier dengan nilai yang sangat tinggi. Kita akan menangani outlier ini dengan metode capping, yaitu membatasi nilai maksimum pada persentil ke-99.



Alasan Penanganan Outliers pada ADR:

Outlier pada 'adr' ditangani dengan capping persentil ke-99 untuk menjaga data tanpa distorsi. Nilai ekstrem bisa membuat distribusi skewed dan memengaruhi model. Batas ini lebih konservatif dari IQR, tetap mempertahankan variabilitas, serta menghindari kesalahan input atau kasus khusus yang tak representatif.

3. **Encoding Categorical Variables**

    Kita akan melakukan encoding terhadap variabel kategorikal dengan metode yang sesuai:

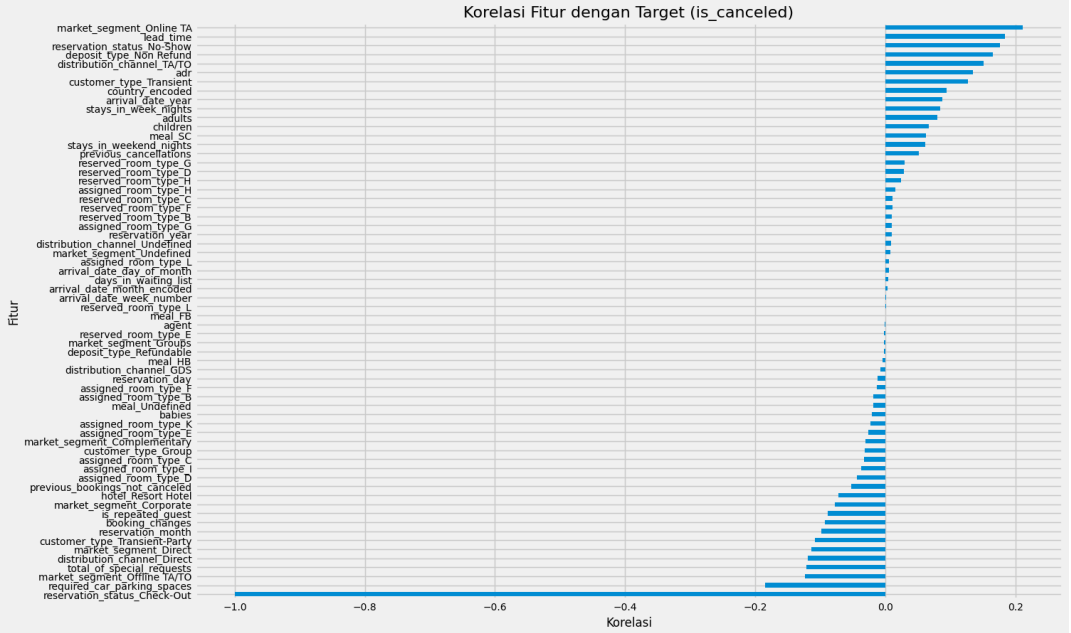
    1. **One-Hot Encoding**: Untuk variabel kategorikal dengan kardinalitas rendah (jumlah kategori sedikit), seperti 'hotel', 'meal', 'market_segment', 'distribution_channel', 'reserved_room_type', 'assigned_room_type', 'deposit_type', 'customer_type', dan 'reservation_status'.

2. **Label Encoding**: Untuk variabel kategorikal dengan kardinalitas tinggi (jumlah kategori banyak), seperti 'country'.
3. **Ordinal Encoding**: Untuk variabel kategorikal yang memiliki urutan, seperti 'arrival_date_month'.

Alasan: Pendekatan ini memungkinkan kita untuk mengubah variabel kategorikal menjadi format numerik yang dapat digunakan dalam model machine learning, dengan mempertimbangkan karakteristik masing-masing variabel.

4. Feature Selection

Feature selection dilakukan dengan memilih fitur yang memiliki korelasi ≥ 0.05 dengan target (is_canceled) untuk meningkatkan performa model dan mengurangi dimensi data.



Berdasarkan analisis korelasi, kita dapat memilih fitur-fitur yang memiliki korelasi yang signifikan dengan target (is_canceled). Kita akan memilih fitur-fitur dengan nilai absolut korelasi di atas threshold tertentu, misalnya 0.05.

```
# Memilih fitur berdasarkan korelasi
threshold = 0.05
selected_features = correlation_with_target.drop('is_canceled').abs()
[correlation_with_target.drop('is_canceled').abs() > threshold].index.tolist()
print(f"Jumlah fitur yang dipilih: {len(selected_features)}")
print(f"Fitur yang dipilih: {selected_features}")
```

Jumlah fitur yang dipilih: 28

```
# Membuat dataset final dengan fitur yang dipilih
X = df_encoded[selected_features]
y = df_encoded['is_canceled']

print(f"Dimensi X: {X.shape}")
print(f"Dimensi y: {y.shape}")
```

Dimensi X: (87230, 28)

Dimensi y: (87230,)

E. Kesimpulan

Pada tugas ini, kelompok kami menganalisis dataset Hotel Booking Demand dari Kaggle, yang berisi 119.390 pemesanan hotel dengan 32 fitur. Tujuan utama adalah memahami proses persiapan data untuk machine learning.

1. Data Description: Dataset mencakup informasi pemesanan hotel, termasuk status pembatalan, jenis hotel, dan harga. Format CSV memudahkan pengolahan dengan Pandas.
2. Data Loading: Dataset dimuat menggunakan read_csv() dari Pandas tanpa tantangan signifikan.
3. Data Understanding: Analisis menunjukkan:
    - City Hotel lebih populer daripada Resort Hotel.
    - 37% pemesanan dibatalkan, dengan lead time tinggi meningkatkan risiko pembatalan.
    - Puncak pemesanan terjadi pada Juli-Agustus.
    - Mayoritas tamu berasal dari Portugal.
4. Data Preparation: Langkah preprocessing meliputi:
    - Menangani missing values (menghapus kolom 'company', mengisi 'agent' dan 'children' dengan 0).
    - Mengatasi outlier pada 'adr' dengan capping persentil ke-99.
    - Encoding variabel kategorikal (One-Hot, Label, dan Ordinal Encoding).
    - Feature selection berdasarkan korelasi ≥ 0.05 dengan target (is_canceled).

Kesimpulannya, persiapan data yang baik, termasuk penanganan missing values, outlier, dan encoding, sangat penting untuk membangun model machine learning yang efektif. Insight dari eksplorasi data membantu dalam pengambilan keputusan selama preprocessing.