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:

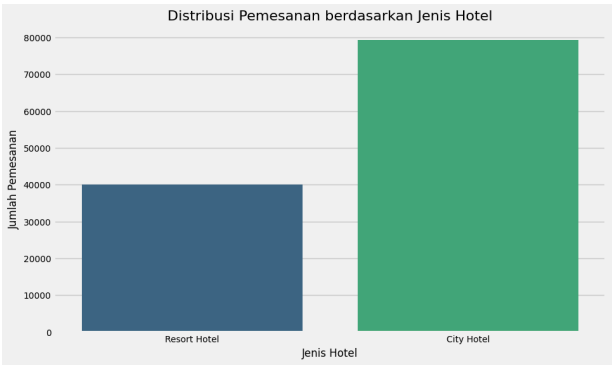
```python
import pandas as pd
import requests
from io import StringIO

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()
```

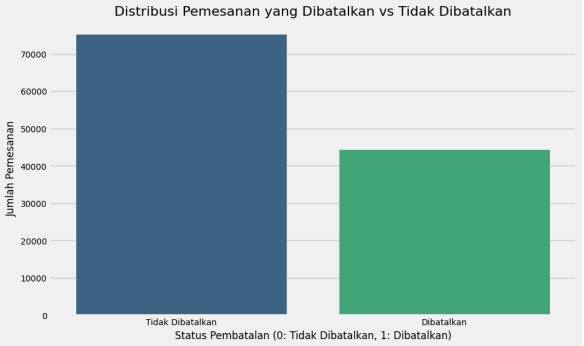
C. Data Understanding

1. Visualisasi Data
   - Distribusi Pemesanan berdasarkan Jenis Hotel



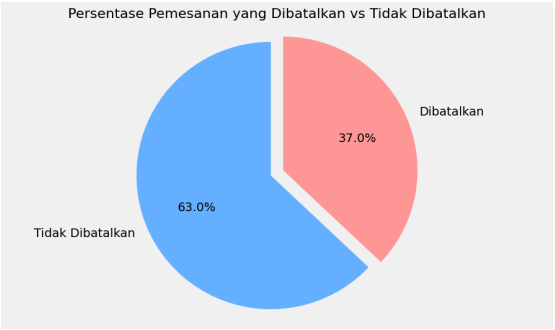
Grafik menunjukkan pemesanan City Hotel lebih tinggi dari Resort Hotel, dengan lebih dari 70.000 vs. sekitar 40.000 pemesanan. Ini menandakan hotel di perkotaan lebih diminati.

- Distribusi Pemesanan yang Dibatalkan vs Tidak Dibatalkan



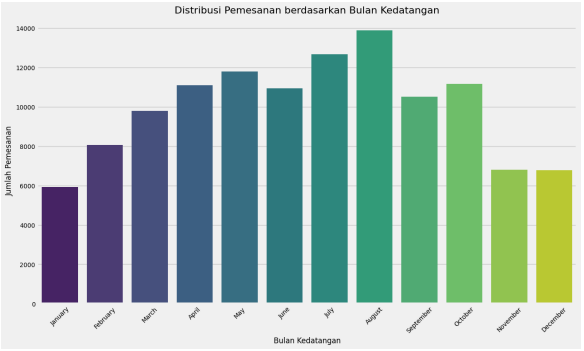
Grafik menunjukkan pemesanan yang tidak dibatalkan lebih banyak (~75.000) dibanding yang dibatalkan (~45.000), meski tingkat pembatalan signifikan, mayoritas pelanggan tetap melanjutkan pesanan.

- Persentase Pemesanan yang Dibatalkan vs Tidak Dibatalkan



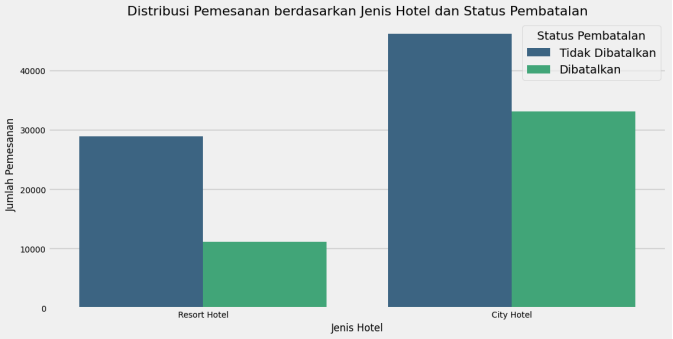
Grafik ini menunjukkan bahwa 37% pemesanan dibatalkan, sementara 63% tetap berlanjut. Tingkat pembatalan yang cukup signifikan ini dapat memengaruhi operasional hotel.

- Distribusi Pemesanan berdasarkan bulan Kedatangan



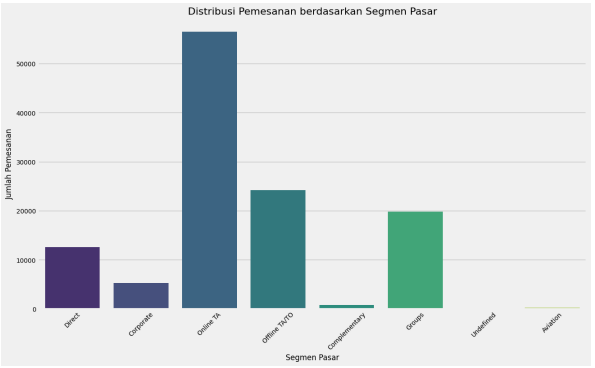
Grafik menunjukkan pemesanan hotel meningkat sejak awal tahun, puncak di Agustus, lalu menurun di akhir tahun, mengindikasikan pola musiman.

- Distribusi Pemesanan berdasarkan Jenis Hotel dan Status Pembatalan



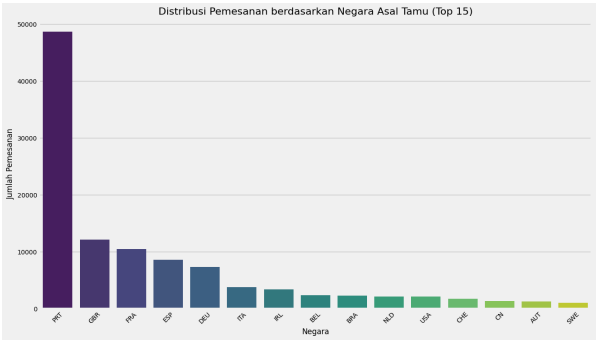
Grafik menunjukkan City Hotel memiliki lebih banyak pemesanan dan tingkat pembatalan lebih tinggi dibanding Resort Hotel, menandakan pemesanannya lebih rentan dibatalkan.

- Distribusi Pemesanan berdasarkan Segmen Pasar



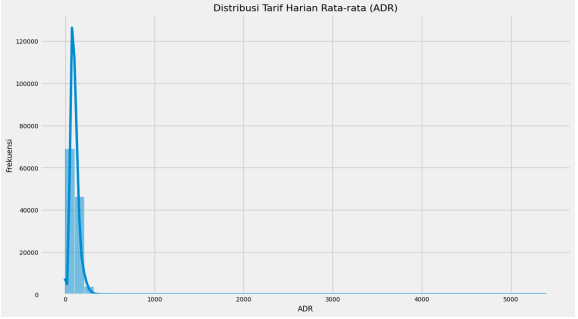
Grafik menunjukkan pemesanan didominasi oleh Online TA, diikuti Offline TA/TO dan Groups. Segmen lainnya jauh lebih kecil, menandakan agen perjalanan online paling banyak digunakan.

- Distribusi Pemesanan berdasarkan Negara Asal Tamu



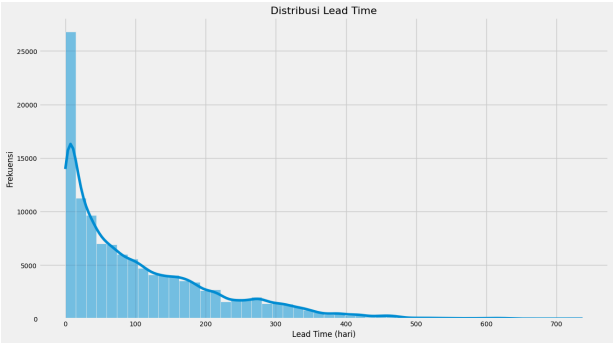
Grafik menunjukkan 15 negara dengan pemesanan tertinggi, dengan satu negara mendominasi. Ini mengindikasikan mayoritas tamu berasal dari negara tersebut, dipengaruhi faktor geografis, kebijakan, atau preferensi wisatawan.

- Distribusi Tarif Harian Rata-rata (ADR)



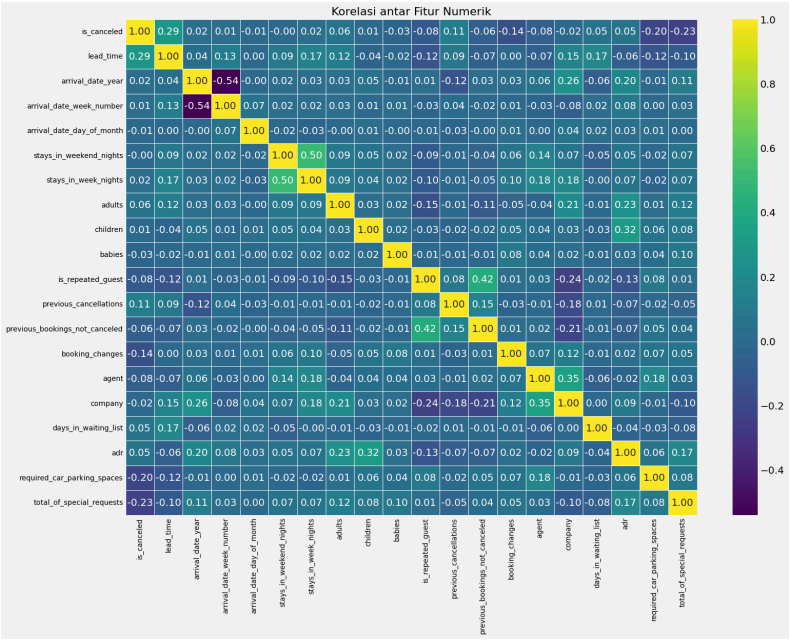
Grafik menunjukkan distribusi ADR skewed ke kanan, dengan mayoritas tarif di bawah 300 dan beberapa outlier tinggi, kemungkinan karena reservasi khusus atau kesalahan data.

- 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.

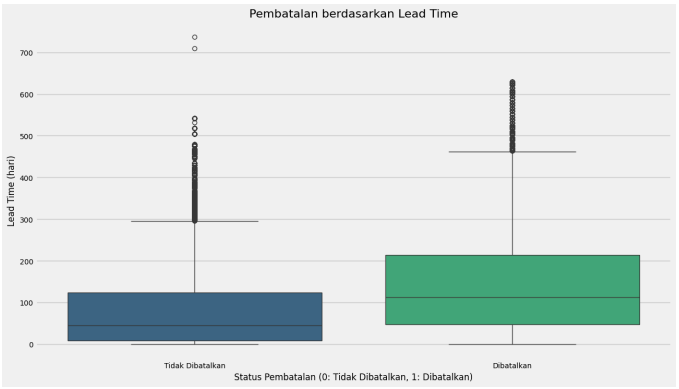
- Korelasi antar Fitur Numerik



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. Tahapan Data Preparation

1. Penanganan Nilai yang Hilang (Handling Missing Values)

   Sebelum penanganan, beberapa kolom dalam dataset memiliki nilai yang hilang:

   - children: 4 nilai hilang
   - country: 488 nilai hilang
   - agent: 16.340 nilai hilang
   - company: 112.593 nilai hilang

   Metode penanganan yang diterapkan:

   - Kolom 'children' diisi dengan nilai 0, dengan asumsi bahwa mayoritas pemesanan tidak melibatkan anak-anak
   - Kolom 'country' diisi dengan 'Unknown', menunjukkan bahwa negara asal tamu tidak diketahui
   - Kolom 'agent' diisi dengan 0, menunjukkan bahwa pemesanan tidak dilakukan melalui agen
   - Kolom 'company' diisi dengan 0, menunjukkan bahwa pemesanan tidak dilakukan melalui perusahaan

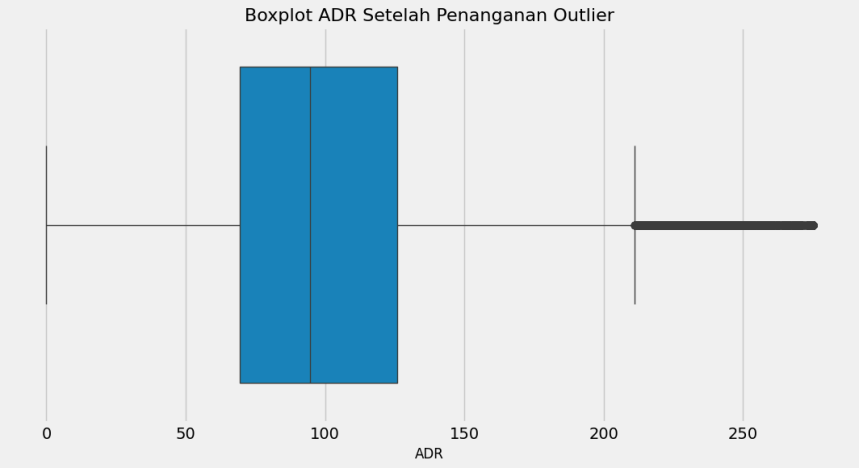
```
Jumlah nilai yang hilang setelah penanganan:
hotel                           0
is_canceled                     0
lead_time                       0
arrival_date_year               0
arrival_date_month              0
arrival_date_week_number        0
arrival_date_day_of_month       0
stays_in_weekend_nights         0
stays_in_week_nights            0
adults                          0
children                        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
company                         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
total_nights                    0
dtype: int64
```

   Setelah penanganan, tidak ada lagi nilai yang hilang dalam dataset.

2. Penanganan Outlier (Handling Outliers)

   Beberapa kolom numerik dalam dataset memiliki outlier yang dapat mempengaruhi analisis:

   - Kolom 'adr' (Average Daily Rate):
     - Nilai negatif tidak masuk akal untuk harga sehingga batas bawah ditentukan pada 0
     - Batas atas ditentukan pada persentil ke-99,5 (275,00)
     - Visualisasi dengan boxplot menunjukkan perbaikan distribusi setelah penanganan
   - Kolom 'adults':
     - Dibatasi maksimum 5 orang dewasa per kamar sebagai batas yang rasional
     - Visualisasi dengan countplot menampilkan distribusi jumlah orang dewasa
   - Kolom 'lead_time':
     - Dibatasi pada persentil ke-99 (444 hari)



     - Visualisasi dengan histogram menunjukkan distribusi lead time setelah penanganan

3. Pengkodean Variabel Kategorikal (Encoding Categorical Variables)

   Kolom-kolom kategorikal dalam dataset dikodekan untuk memungkinkan penggunaan dalam analisis lanjutan:

   - Label Encoding diterapkan pada:
     - hotel: Resort Hotel (0), City Hotel (1)
     - meal: BB (0), FB (1), HB (2), SC (3), Undefined (4)
     - market_segment: Direct (0), Corporate (1), Online TA (2), dll.
     - distribution_channel: Direct (0), Corporate (1), TA/TO (2), dll.
     - reserved_room_type & assigned_room_type: A (0), B (1), C (2), dll.
     - deposit_type: No Deposit (0), Refundable (1), Non Refund (2)

    - customer_type: Transient (0), Contract (1), Transient-Party (2), Group (3)
    - reservation_status: Check-Out (0), Canceled (1), No-Show (2)
    - country: Menggunakan indeks unik untuk 178 nilai unik
    - arrival_date_month: January (1), February (2), March (3), dll.
- Ekstraksi Waktu dari reservation_status_date:
    - Dikonversi menjadi format datetime
    - Diekstrak menjadi kolom status_year, status_month, dan status_day

4. Rekayasa Fitur (Feature Engineering)

   Beberapa fitur baru dibuat untuk meningkatkan kualitas analisis:

   - total_nights: Jumlah total malam menginap (stays_in_weekend_nights + stays_in_week_nights)
   - total_guests: Jumlah total tamu (adults + children + babies)
   - is_high_season: Menandai pemesanan pada musim tinggi (July, August, December)
   - lead_time_category: Mengkategorikan lead time menjadi short (0-30 hari), medium (31-90 hari), dan long (>90 hari)
   - is_weekend_arrival: Menandai kedatangan pada akhir pekan
   - has_special_requests: Menandai pemesanan dengan permintaan khusus
   - is_repeated_guest_with_previous_cancellations: Menandai tamu berulang yang pernah membatalkan pemesanan
   - price_per_person: Harga per orang (adr / total_guests)
   - room_assignment_match: Menandai apakah tipe kamar yang dipesan sama dengan yang diberikan

5. Seleksi Fitur (Feature Selection)

   Seleksi fitur dilakukan untuk memilih fitur yang paling relevan untuk analisis:

   1. Kolom yang redundan atau sudah dienkode dihapus
   2. Teknik ANOVA F-value digunakan untuk memilih 15 fitur teratas yang paling berpengaruh terhadap variabel target is_canceled
   3. Fitur-fitur teratas berdasarkan skor:
      - reservation_status_encoded (skor tertinggi: 2.99e+06)
      - deposit_type_encoded (3.58e+04)
      - lead_time (1.13e+04)
      - has_special_requests (8.99e+03)
      - room_assignment_match (7.81e+03)
      - market_segment_encoded (7.19e+03)
      - total_of_special_requests (6.96e+03)
      - required_car_parking_spaces (4.74e+03)
      - assigned_room_type_encoded (3.82e+03)
      - distribution_channel_encoded (3.54e+03)

    - booking_changes (2.54e+03)
    - hotel_encoded (2.27e+03)
    - customer_type_encoded (2.24e+03)
    - previous_cancellations (1.47e+03)
    - country_encoded (1.22e+03)
4. Dataset final terdiri dari 16 kolom (15 fitur terpilih + variabel target) dengan 119.390 baris data

Insight dari Preprocessing

- Fitur Penting: Faktor utama dalam prediksi pembatalan adalah *lead_time*, *deposit_type_encoded*, *total_of_special_requests*, dan *previous_cancellations*. Ini menunjukkan bahwa jarak waktu sebelum kedatangan, jenis deposit, jumlah permintaan khusus, dan riwayat pembatalan sebelumnya berpengaruh signifikan.
- Pola Musiman: *is_high_season* dan *arrival_date_month_encoded* menunjukkan adanya pola musiman dalam pembatalan, dengan karakteristik berbeda antara musim tinggi dan rendah.
- Karakteristik Tamu: Jumlah tamu (*total_guests*, *adults*) serta status tamu berulang (*is_repeated_guest*) juga berperan, menunjukkan bahwa tamu lama dan baru memiliki pola pembatalan berbeda.
- Harga dan Nilai: *adr* dan *price_per_person* berpengaruh terhadap keputusan pembatalan, di mana harga yang lebih tinggi bisa memiliki pola pembatalan berbeda dari harga lebih rendah.

E. Kesimpulan

Dataset Hotel Booking Demand dari Kaggle berisi 119.390 data pemesanan hotel dari tahun 2015 hingga 2017 dengan 32 fitur. Mayoritas pemesanan dilakukan di City Hotel dibandingkan Resort Hotel. Tingkat pembatalan mencapai 37%, dengan pola musiman menunjukkan puncak pemesanan pada Juli-Agustus. Segmen Online Travel Agent (OTA) mendominasi pemesanan, sementara sebagian besar tamu berasal dari Portugal dan negara Eropa lainnya.

Dalam tahap data preparation, nilai yang hilang telah ditangani dengan pengisian nilai default yang sesuai, sementara outlier dalam fitur seperti tarif harian dan jumlah tamu telah dikendalikan. Variabel kategorikal dikodekan untuk analisis lebih lanjut, dan beberapa fitur baru dibuat untuk meningkatkan wawasan, seperti total malam menginap, total tamu, dan kategori lead time.

Seleksi fitur dilakukan untuk memilih fitur paling relevan terhadap status pembatalan pemesanan. Fitur yang memiliki pengaruh kuat termasuk reservation_status, deposit_type, lead_time, dan has_special_requests. Dataset akhir terdiri dari 16 kolom dengan fitur terpilih yang siap digunakan dalam analisis lanjutan atau pemodelan prediktif.