Analisis Data Eksploratif (EDA) pada Dataset Hotel Booking Demand

Fadhli Ilham Nafi'an Yuswono, Karina Aulia Sari, Dosen Pembimbing 2
Teknik Informatika, Institut Teknologi Nasional Malang
Jalan Raya Karanglo km 2 Malang, Indonesia
2318001@scholar.itn.ac.id

ABSTRAK

Analisis data eksploratif (EDA) merupakan tahap penting dalam proses ilmu data untuk mengetahui ciri-ciri utama dari sebuah *dataset*. Penelitian ini melaksanakan EDA pada *dataset* "Hotel Booking Demand" yang memiliki lebih dari 119.000 entri pemesanan untuk *City Hotel* dan *Resort Hotel*. Analisis ini bertujuan untuk mengenali pola pemesanan, mengkaji faktor-faktor yang berperan dalam pembatalan pemesanan, dan memperoleh wawasan tambahan yang dapat diimplementasikan. Metodologi yang diterapkan mencakup pembersihan data, analisis univariat, bivariat, serta multivariat dengan memanfaatkan bahasa pemrograman Python dan pustaka Pandas, Matplotlib, serta Seaborn. Hasil utama mengindikasikan bahwa (1) *City Hotel* mencatat *volume* pemesanan dan tingkat pembatalan yang lebih tinggi dibandingkan *Resort Hotel*; (2) Waktu tunggu (*lead time*) yang lebih panjang berkaitan positif dengan kemungkinan terjadinya pembatalan; dan (3) Puncak pemesanan terjadi pada bulan Agustus, sedangkan harga kamar rata-rata (ADR) juga berfluktuasi secara musiman.

Kata kunci : *Analisis Data Eksploratif, Data Mining, Pembelajaran Mesin, Data Analis*

ABSTRACT

Exploratory data analysis (EDA) is an important stage in the data science process to determine the main characteristics of a dataset. This study conducted EDA on the 'Hotel Booking Demand' dataset, which has more than 119,000 booking entries for City Hotels and Resort Hotels. This analysis aims to recognise booking patterns, examine factors that play a role in booking cancellations, and gain additional insights that can be implemented. The methodology applied includes data cleaning, univariate, bivariate, and multivariate analysis using the Python programming language and the Pandas, Matplotlib, and Seaborn libraries. The main results indicate that (1) City Hotel recorded higher booking volumes and cancellation rates than Resort Hotel; (2) longer lead times are positively associated with the likelihood of cancellations; and (3) booking peaks occur in August, while average daily rates (ADR) also fluctuate seasonally.

Keywords : *Exploratory Data Analysis, Data Mining, Machine Learning, Data Analyst*

1. PENDAHULUAN

Industri perhotelan adalah sektor yang sangat kompetitif dan dinamis, di mana manajemen pendapatan (*revenue management*) menjadi faktor kunci keberhasilan. Salah satu tantangan utama yang dihadapi sektor ini adalah tingginya angka pembatalan pesanan. Pembatalan tidak hanya mengakibatkan hilangnya pendapatan langsung tetapi juga menyulitkan pengaturan inventaris kamar dan distribusi sumber daya. Untuk menyelesaikan masalah ini, manajemen hotel harus memahami pola perilaku tamu dan mengidentifikasi faktor-faktor yang berkontribusi pada keputusan pembatalan.

Memahami data pemesanan masa lalu merupakan langkah pertama yang penting. Dataset "Hotel Booking Demand" menyajikan catatan mendetail dari ratusan ribu reservasi yang mencakup berbagai atribut, mulai dari waktu pemesanan, durasi menginap, jenis pelanggan, hingga status akhir dari pemesanan. Sebelum membuat model prediktif yang rumit seperti *machine learning*, penting untuk melakukan Analisis Data Eksploratif (EDA) terlebih dahulu.

EDA merupakan tahapan eksplorasi awal terhadap data untuk mengidentifikasi pola, mendeteksi *outlier*, dan merangkum karakteristik utama, biasanya dengan memanfaatkan visualisasi data. Studi ini bertujuan untuk menerapkan metode EDA pada kumpulan data "Permintaan Pemesanan Hotel". Fokus analisis bertujuan untuk menjawab pertanyaan utama seperti: Apa saja ciri-ciri umum dari pemesanan? Apa yang menjadi alasan utama membedakan antara pemesanan yang dibatalkan dan yang tetap? Bagaimana pola musiman mempengaruhi jumlah pemesanan dan harga, dll.

2. METODE

Metodologi penelitian ini menjelaskan langkah-langkah yang diambil dalam melakukan Analisis Data Eksploratif (EDA) pada dataset "Hotel Booking Demand".

Gambar 2.1 Alur Tahapan EDA

2.1. Sumber Data

Data yang digunakan adalah dataset "Hotel Booking Demand" yang bersumber dari platform Kaggle. Dataset ini terdiri dari 119.390 baris dan 32 kolom sebelum proses pembersihan, yang mencakup data pemesanan untuk City Hotel dan Resort Hotel.

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

2.2. Library dan Bahasa Pemrograman

Analisis dilakukan menggunakan bahasa pemrograman Python dalam lingkungan Google Colaboratory. *Library* utama yang digunakan meliputi:

Pandas : untuk pemuatan, manipulasi, dan pembersihan data

Matplotlib & Seaborn : untuk visualisasi data, ex: analisis univariat, bivariat, dan multivariat

Missingno : untuk memvisualisasikan data yang hilang/*missing values*.

2.3. Tahapan Analisis

Agar analisis dapat dilakukan secara menyeluruh dan teratur, seluruh proses Analisis Data Eksploratif (EDA) dalam studi ini dibagi menjadi empat langkah utama. Langkah-langkah ini dirancang untuk dilaksanakan secara teratur dan berurutan, dimulai dari pengumpulan data hingga analisis multivariat, sebagai berikut :

2.3.1. Pemuatan dan Inspeksi Data

Memuat *dataset* ke dalam *DataFrame* Pandas dan melakukan pemeriksaan awal menggunakan df.info(), df.head(), dan df.describe() untuk memahami struktur, tipe data, dan statistik deskriptif awal.

```
Dataset berhasil dimuat. Jumlah baris: 119390, Jumlah kolom: 32
          hotel  is_canceled  lead_time  arrival_date_year arrival_date_month  \
0  Resort Hotel            0        342               2015               July
1  Resort Hotel            0        737               2015               July
2  Resort Hotel            0          7               2015               July
3  Resort Hotel            0         13               2015               July
4  Resort Hotel            0         14               2015               July

   arrival_date_week_number  arrival_date_day_of_month  \
0                        27                          1
1                        27                          1
2                        27                          1
3                        27                          1
4                        27                          1

   stays_in_weekend_nights  stays_in_week_nights  adults  ...  deposit_type  \
0                        0                     0       2  ...    No Deposit
1                        0                     0       2  ...    No Deposit
2                        0                     1       1  ...    No Deposit
3                        0                     1       1  ...    No Deposit
4                        0                     2       2  ...    No Deposit

   agent company days_in_waiting_list customer_type   adr  \
0    NaN     NaN                    0     Transient   0.0
1    NaN     NaN                    0     Transient   0.0
2    NaN     NaN                    0     Transient  75.0
3  304.0     NaN                    0     Transient  75.0
4  240.0     NaN                    0     Transient  98.0

   required_car_parking_spaces  total_of_special_requests  reservation_status  \
0                            0                          0           Check-Out
1                            0                          0           Check-Out
2                            0                          0           Check-Out
3                            0                          0           Check-Out
4                            0                          1           Check-Out

  reservation_status_date
0              2015-07-01
1              2015-07-01
2              2015-07-02
3              2015-07-02
4              2015-07-03

[5 rows x 32 columns]
```

Gambar 2.2 Tampilan df.head()

```
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         118902 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  object
dtypes: float64(4), int64(16), object(12)
```

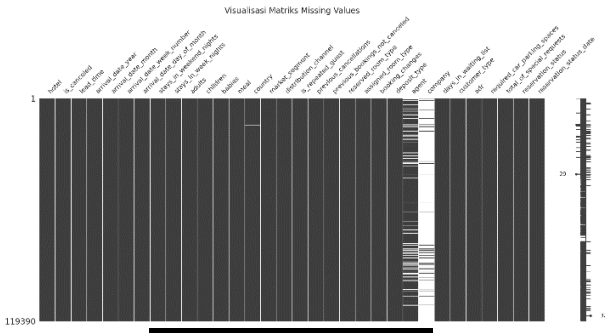
Gambar 2.3 Tampilan df.info()

```
          is_canceled      lead_time  arrival_date_year  \
count  119390.000000  119390.000000      119390.000000
mean        0.370416     104.011416        2016.156554
std         0.482918     106.863097           0.707476
min         0.000000       0.000000        2015.000000
25%         0.000000      18.000000        2016.000000
50%         0.000000      69.000000        2016.000000
75%         1.000000     160.000000        2017.000000
max         1.000000     737.000000        2017.000000

       arrival_date_week_number  arrival_date_day_of_month  \
count             119390.000000              119390.000000
mean                  27.165173                  15.798241
std                   13.605138                   8.780829
min                    1.000000                   1.000000
25%                   16.000000                   8.000000
50%                   28.000000                  16.000000
75%                   38.000000                  23.000000
max                   53.000000                  31.000000

       stays_in_weekend_nights  stays_in_week_nights        adults  \
count            119390.000000         119390.000000  119390.000000
mean                  0.927599              2.500302       1.856403
std                   0.998613              1.908286       0.579261
min                   0.000000              0.000000       0.000000
25%                   0.000000              1.000000       2.000000
50%                   1.000000              2.000000       2.000000
75%                   2.000000              3.000000       2.000000
max                  19.000000             50.000000      55.000000

            children         babies  is_repeated_guest  \
count  119386.000000  119390.000000      119390.000000
mean        0.103890       0.007949           0.031912
std         0.398561       0.097436           0.175767
min         0.000000       0.000000           0.000000
25%         0.000000       0.000000           0.000000
50%         0.000000       0.000000           0.000000
75%         0.000000       0.000000           0.000000
max        10.000000      10.000000           1.000000

       previous_cancellations  previous_bookings_not_canceled  \
count           119390.000000                   119390.000000
mean                0.087118                        0.137097
std                 0.844336                        1.497437
min                 0.000000                        0.000000
25%                 0.000000                        0.000000
50%                 0.000000                        0.000000
75%                 0.000000                        0.000000
max                26.000000                       72.000000

       booking_changes          agent      company  days_in_waiting_list  \
count    119390.000000  103050.000000  6797.000000         119390.000000
mean          0.221124      86.693382   189.266735              2.321149
std           0.652306     110.774548   131.655015             17.594721
min           0.000000       1.000000     6.000000              0.000000
25%           0.000000       9.000000    62.000000              0.000000
50%           0.000000      14.000000   179.000000              0.000000
75%           0.000000     229.000000   270.000000              0.000000
max          21.000000     535.000000   543.000000            391.000000

                 adr  required_car_parking_spaces  total_of_special_requests
count  119390.000000                119390.000000              119390.000000
mean      101.831122                     0.062518                   0.571363
std        50.535790                     0.245291                   0.792798
min        -6.380000                     0.000000                   0.000000
25%        69.290000                     0.000000                   0.000000
50%        94.575000                     0.000000                   0.000000
75%       126.000000                     0.000000                   1.000000
max      5400.000000                     8.000000                   5.000000
```

Gambar 2.4 Tampilan df.describe()

2.3.2. Data Cleaning

Menganalisis kolom dengan data hilang seperti *company*, *agent*, *country*, *children*. Kolom *company* lebih dari 90% hilang, maka dihapus. Kolom agent diisi dengan 0 dengan asumsi tidak pakai agen. Kolom *country* dan children diisi dengan modus dan 0. Serta Mengidentifikasi dan menghapus baris data yang tidak logis, seperti pemesanan dengan jumlah total tamu *adults* + *children* + *babies* = 0




```
company                         112593
agent                            16340
country                            488
children                             4
arrival_date_month                   0
arrival_date_week_number             0
hotel                                0
is_canceled                          0
stays_in_weekend_nights              0
arrival_date_day_of_month            0
adults                               0
stays_in_week_nights                 0
babies                               0
meal                                 0
lead_time                            0
arrival_date_year                    0
distribution_channel                 0
market_segment                       0
previous_bookings_not_canceled       0
is_repeated_guest                    0
reserved_room_type                   0
assigned_room_type                   0
booking_changes                      0
previous_cancellations               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
```

Gambar 2.5 Tampilan *Missing Values*

```
--- Missing Values SETELAH Cleaning ---
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
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
```

Gambar 2.6 Tampilan Hasil *Data Cleansing*

2.3.3. Analisis Univariat

Menganalisis distribusi dari satu variabel tunggal untuk memahami karakteristiknya. Misalnya *countplot* untuk variabel kategorikal (is_canceled, hotel) dan histogram untuk variabel numerik (lead_time).

2.3.4. Analisis Bivariat dan Multivariat

Menganalisis hubungan antara dua atau lebih variabel untuk menemukan pola. Misalnya *countplot* dengan parameter hue untuk Hotel & Pembatalan, *boxplot* untuk *Lead Time* & Pembatalan, *lineplot* untuk Tren ADR per bulan, dan *heatmap* korelasi untuk semua variabel numerik.

3. HASIL DAN PEMBAHASAN

Pada bagian ini, akan dipaparkan secara mendalam temuan-temuan esensial serta wawasan / *insight* yang berhasil diperoleh dari serangkaian proses analisis data yang telah dilakukan. Pembahasan akan mencakup hasil dari pembersihan data, temuan dari analisis univariat, hingga pola-pola signifikan yang terungkap melalui analisis bivariat dan multivariat.

3.1 Input dan Data Cleansing

Dari pemeriksaan awal, ditemukan empat kolom dengan *missing values* signifikan. Kolom *company* dihapus karena 94% datanya kosong. Kolom *agent* diisi nilai 0 untuk merepresentasikan pemesanan langsung. Sejumlah 180 baris data ditemukan tidak valid karena tidak memiliki tamu (0 dewasa, 0 anak, 0 bayi) dan kemudian dihapus. Dataset akhir yang bersih berisi 119.210 baris data yang siap untuk dianalisis.

```
Dataset berhasil dimuat. Jumlah baris: 119390, Jumlah kolom: 32
```

Gambar 3.1 Tampilan Data Awal

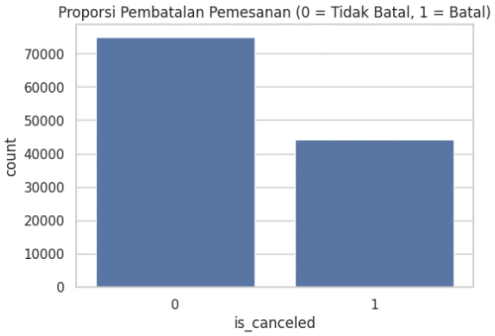
```
Jumlah baris setelah menghapus data 0 tamu: 119210
```

Gambar 3.2 Tampilan Setelah *Data Cleansing*

3.2 Analisis Univariat

Analisis pada level univariat difokuskan pada pengujian distribusi dan karakteristik dari setiap variabel kunci secara individual. Proses ini penting untuk memahami komposisi dasar dari dataset sebelum melangkah ke analisis yang lebih kompleks. Beberapa temuan fundamental mengenai dataset ini ditunjukkan sebagai berikut :

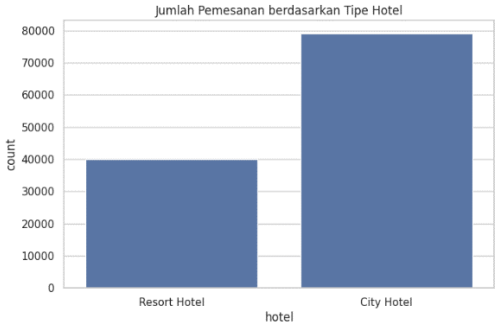
3.2.1 Proporsi Pembatalan



Gambar 3.3 Tampilan Presentase Bembatalan

Dari total pemesanan, 37,04% berakhir dengan pembatalan (*is_canceled* = 1), sementara 62,96% sisanya dikonfirmasi. Ini menunjukkan bahwa pembatalan adalah masalah signifikan.

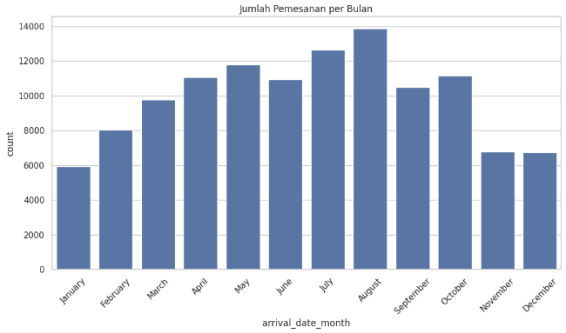
3.2.2 Tipe Hotel



Gambar 3.4 Tampilan Presentase Tipe Hotel

City Hotel (66,4%) jauh lebih mendominasi *dataset* dibandingkan *Resort Hotel* (33,6%)

3.2.3 Pola Musiman



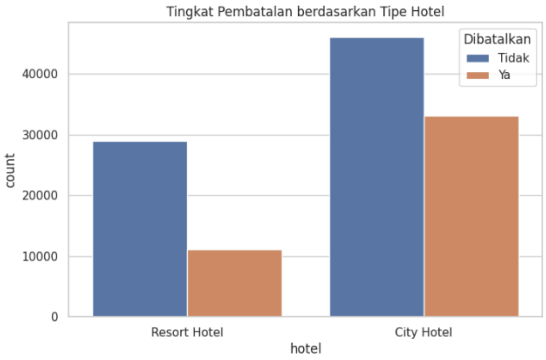
Gambar 3.5 Tampilan Kedatangan Bulanan

Analisis bulan kedatangan (*arrival_date_month*) menunjukkan puncak musim pemesanan terjadi pada bulan Agustus, diikuti oleh Juli dan Mei. Bulan dengan pemesanan terendah adalah Januari dan November.

3.3 Analisis Bivariat dan Multivariat

Setelah mengetahui ciri-ciri tiap variabel, analisis dilanjutkan untuk mengeksplorasi hubungan, pola, dan korelasi yang ada *antar* variabel (bivariat dan multivariat). Eksplorasi ini berhasil mengungkap pemahaman yang lebih mendalam mengenai faktor-faktor yang saling mempengaruhi, terutama yang berkaitan dengan status pembatalan hotel.

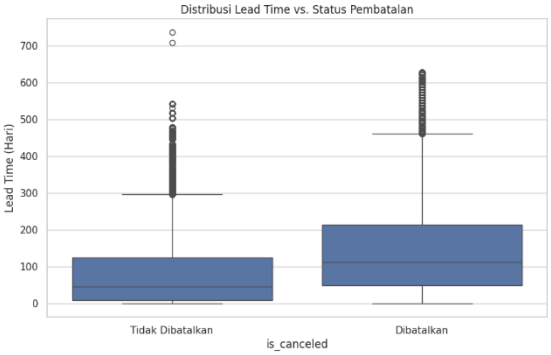
3.3.1 Tipe Hotel & Pembatalan



Gambar 3.6 Tampilan Tingkat Pembatalan Berdasarkan Tipe Hotel

Ditemukan bahwa *City Hotel* tidak hanya memiliki volume pemesanan lebih tinggi, tetapi juga tingkat pembatalan yang secara proporsional lebih tinggi (sekitar 41,7%) dibandingkan dengan *Resort Hotel* (27,8%).

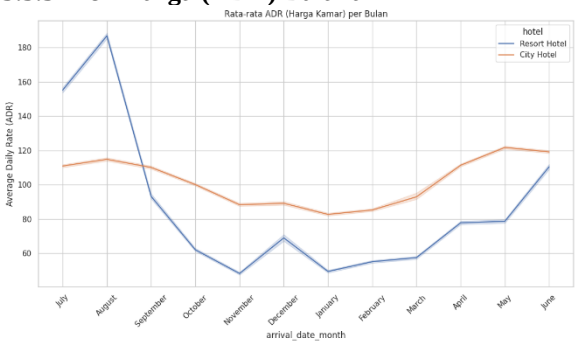
3.3.2 Waktu Tunggu (Lead Time) & Pembatalan



Gambar 3.7 Tampilan Distribusi *Lead Time* & Pembatalan

Boxplot menunjukkan hubungan yang jelas dimana pemesanan yang dibatalkan memiliki median *lead time*/waktu tunggu jauh lebih panjang. Tamu yang memesan jauh-jauh hari lebih cenderung untuk membatalkan dibandingkan tamu yang memesan mendekati tanggal kedatangan.

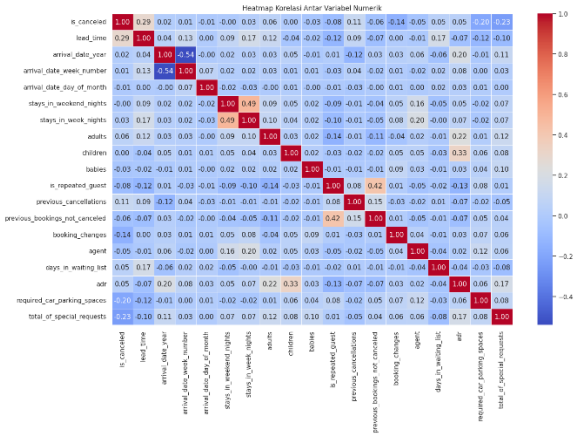
3.3.3 Tren Harga (ADR) Bulanan



Gambar 3.8 Tampilan ADR Bulanan

Harga rata-rata harian (ADR) bervariasi secara musiman. *Resort Hotel* menunjukkan lonjakan harga yang signifikan di musim puncak (Juli-Agustus), sedangkan *City Hotel* memiliki harga yang relatif lebih konsisten/stabil sepanjang tahun.

3.3.4 Analisis Korelasi



Gambar 3.9 Tampilan Heatmap

Heatmap korelasi antar variabel numerik menunjukkan korelasi positif sedang (0.29) antara *lead_time* dan *is_canceled*, mengonfirmasi temuan boxplot. Kemudian korelasi negatif sedang (-0.23) antara *total_of_special_requests* dan *is_canceled*. Ini

adalah temuan penting, dimana tamu yang memiliki permintaan Khusus, misal *connecting room, high floor* jauh lebih kecil kemungkinannya untuk membatalkan.

4. KESIMPULAN

Analisis Data Eksploratif (EDA) pada dataset "Hotel Booking Demand" telah berhasil menemukan berbagai pola dan wawasan penting. Hasil utama menunjukkan bahwa tingkat pembatalan secara keseluruhan sangat tinggi (37,04%), dengan angka yang jauh lebih tinggi di City Hotel (41,7%) dibandingkan Resort Hotel.

Faktor prediktif terkuat untuk pembatalan adalah *lead time* yang panjang; semakin lama jeda antara pemesanan dan *check-in*, semakin tinggi risiko pembatalan. Sebaliknya, keterlibatan tamu, yang ditunjukkan oleh *total_of_special_requests*, secara signifikan mengurangi risiko pembatalan. Secara musiman, pemesanan dan harga kamar memuncak di bulan-bulan musim panas, terutama Agustus.

Temuan ini dapat digunakan oleh manajemen hotel untuk merancang strategi mitigasi, seperti menerapkan kebijakan deposit yang lebih ketat pada pemesanan dengan waktu tunggu panjang atau melakukan pendekatan proaktif kepada tamu yang tidak memiliki permintaan khusus.

5. TINJAUAN PUSTAKA

Analisis Data Eksploratif atau *Exploratory Data Analysis* (EDA) adalah pendekatan dasar dalam analisis data yang bertujuan untuk merangkum sifat-sifat utama dari suatu dataset, biasanya dengan menggunakan teknik visual [1]. Tujuan utama EDA bukanlah untuk menguji hipotesis secara formal, tetapi untuk melakukan penyelidikan awal guna menemukan pola yang tersembunyi, mengidentifikasi anomali atau *outlier*, serta memahami hubungan antara variabel. Langkah ini menjadi sangat penting sebelum beralih ke tahap pemodelan yang lebih kompleks [2].

Di sektor perhotelan, salah satu tantangan operasional utama adalah pengelolaan pendapatan (*revenue management*), di mana tingkat pembatalan pemesanan (*cancellation rate*) berperan penting dalam memengaruhi profitabilitas [3]. **Antonio, de Almeida, dan Nunes (2019)** menyediakan dataset publik "Hotel Booking Demand" untuk mendukung analisis di bidang ini, yang mencakup data pemesanan nyata dari *City Hotel* dan *Resort Hotel* di Portugal [4]. Dataset ini sangat beragam dan lengkap, mencakup berbagai variabel seperti waktu tunggu (*lead time*), demografi pengunjung, jenis pemesanan, hingga status akhir pemesanan, sehingga menjadi referensi terkenal bagi peneliti untuk memahami faktor-faktor yang memengaruhi keputusan pembatalan.

Beragam analisis sebelumnya yang menggunakan dataset ini secara konsisten menemukan sejumlah wawasan penting. Waktu tunggu atau lead time yang merupakan selang hari antara tanggal pemesanan dan tanggal kedatangan sering kali diidentifikasi sebagai salah satu indikator pembatalan yang paling penting [5]. Temuan lain yang signifikan adalah terdapat hubungan negatif antara total permintaan khusus dengan tingkat pembatalan, yang menunjukkan bahwa pelanggan yang lebih terlibat cenderung tidak membatalkan pesanan mereka [6].

6. DAFTAR PUSTAKA

Tukey, J. W. (1977). Exploratory data analysis. Reading/Addison-Wesley.

Chatfield, C. (1985). The initial examination of data. Journal of the Royal Statistical Society: Series A (General), 148(3), 214-231.

Antonio, N., de Almeida, A., & Nunes, L. (2019). Hotel booking demand datasets. Data in brief, 22, 41-49.

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