Digital Transformation Technology (Digitech) | e-ISSN : 9999-9999
Volume 1, **Number** 1, June 2021
https://doi.org/10.47709/briliance.vxix.xxxx

Analisis Data Eksploratif (EDA) pada Dataset Hotel Booking Demand

Fadhli Ilham Nafi'an Yuswono[1*], Karina Aulia Sari[2]
[1]Teknik Informatika, Institut Teknologi Nasional Malang, Indonesia, [2]Teknik Informatika, Institut Teknologi Nasional Malang, Indonesia
2318001@student.itn.ac.id, karina.auliasari86@gmail.com

open access
Histori Artikel:
Diajukan: 24 Oktober 2025
Disetujui: 24 Oktober 2025
Dipublikasi: 25 Oktober 2025
Kata Kunci:
Analisis Data Eksploratif; Data Mining; *Machine Learning*; Data Analis; Google Colab

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.

PENDAHULUAN

Industri perhotelan merupakan 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 tingkat pembatalan pesanan. Pembatalan tidak hanya mengakibatkan hilangnya pendapatan langsung, tetapi juga menyulitkan pengelolaan inventaris kamar dan distribusi sumber daya. Untuk mengatasi masalah ini, manajemen hotel perlu memahami pola perilaku tamu dan mengidentifikasi faktor-faktor yang mempengaruhi pada keputusan pembatalan.

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

EDA merupakan langkah awal dalam eksplorasi 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.

STUDI LITERATUR

Analisis Data Eksploratif atau Exploratory Data Analysis (EDA) merupakan pendekatan fundamental dalam analisis data yang bertujuan untuk merangkum karakteristik utama dari sebuah dataset, biasanya dengan menggunakan teknik visualisasi. Tujuan utama EDA bukanlah untuk menguji hipotesis secara formal, tetapi untuk melakukan ekplorasi 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.

Di industri perhotelan, salah satu tantangan operasional utama adalah pengelolaan pendapatan (*revenue management*), di mana tingkat pembatalan pemesanan (*cancellation rate*) berperan penting dalam memengaruhi profitabilitas. 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. *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.

Berbagai 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. 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.

METODE

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

Gambar 1 Alur Tahapan EDA

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

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

Library dan Bahasa Pemrograman

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

1. Pandas : untuk pemuatan, manipulasi, dan pembersihan data
2. Matplotlib & Seaborn : untuk visualisasi data, ex: analisis univariat, bivariat, dan multivariat.
3. Missingno : untuk memvisualisasikan data yang hilang/missing values.

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.

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.

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

3. **Analisi 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).

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.

HASIL

Input dan Data Cleansing

Dataset awal berhasil dimuat dengan 119.390 baris dan 32 kolom

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

Gambar 2 tampilan membaca *dataset*

Tabel 1 tampilan df.head() untuk 5 baris awal data

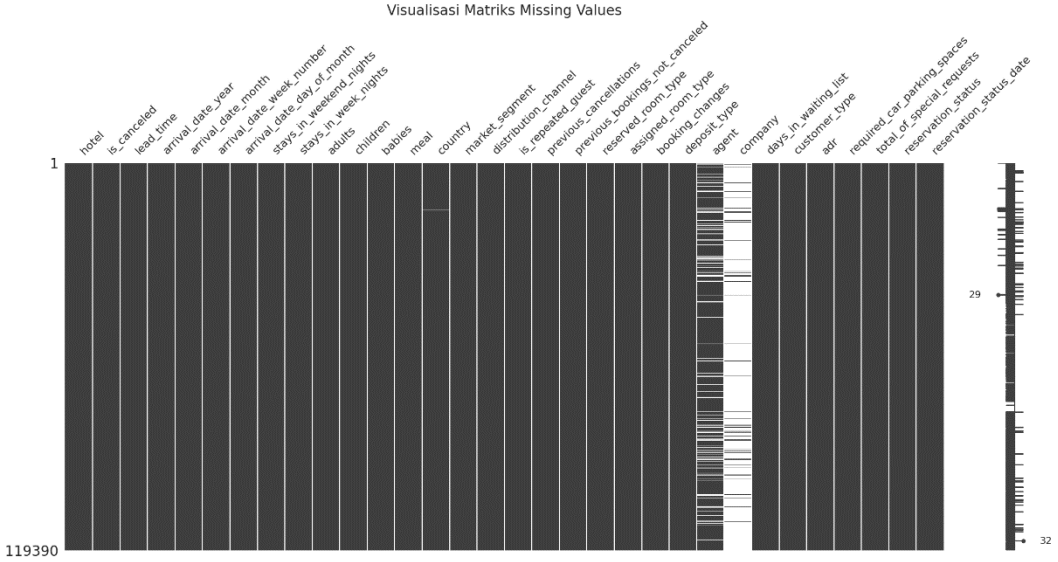
hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	...	deposit_type	agent	company	days_in_waiting_list	customer_type	adr	required_car_parking_spaces	total_of_special_requests	reservation_status	reservation_status_date
Resort Hotel	0	342	2015	July	27	1	0	0	2	...	No Deposit	NaN	NaN	0	Transient	00.00	0	0	Check-Out	01/07/2015
Resort Hotel	0	737	2015	July	27	1	0	0	2	...	No Deposit	NaN	NaN	0	Transient	00.00	0	0	Check-Out	01/07/2015
Resort Hotel	0	7	2015	July	27	1	0	1	1	...	No Deposit	NaN	NaN	0	Transient	75.00.00	0	0	Check-Out	02/07/2015
Resort Hotel	0	13	2015	July	27	1	0	1	1	...	No Deposit	304.00.00	NaN	0	Transient	75.00.00	0	0	Check-Out	02/07/2015
Resort Hotel	0	14	2015	July	27	1	0	2	2	...	No Deposit	240.00.00	NaN	0	Transient	98.00.00	0	1	Check-Out	03/07/2015

```
 #   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)
```

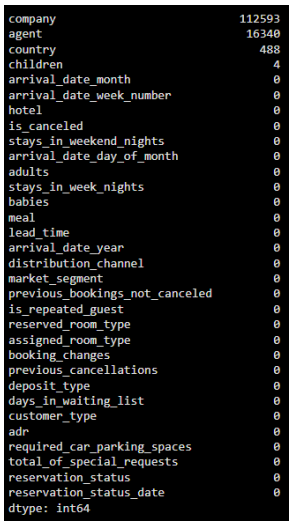
Gambar 3 tampilan df.info()

Tabel 2 tampilan df.describe()

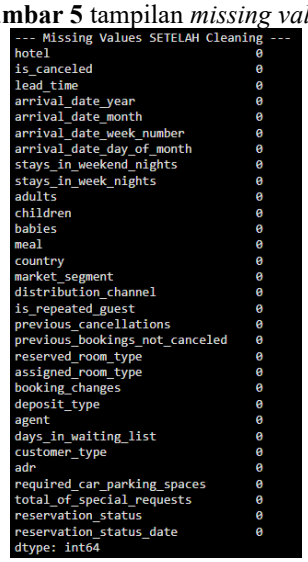
	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled	booking_changes	agent	days_in_waiting_list	adr	required_car_parking_spaces	total_of_special_requests
count	119.210.000.000																		
mean	0.370766	104.109.227	2.016.156.472	27.165.376	15.798.717	0.927053	2.499.195	1.859.206	0.104043	0.007961	0.031499	0.087191	0.137094	0.218799	74.889.078	2.321.215	101.969.092	0.062553	0.571504
std	0.483012	106.875.450	0.707485	13.601.107	8.781.070	0.995117	1.897.106	0.575186	0.398836	0.097509	0.174663	0.844918	1.498.137	0.638504	107.168.884	17.598.002	50.434.007	0.245360	0.792876
min	0.000000	0.000000	2.015.000.000	1.000.000	1.000.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-6.380.000	0.000000	0.000000
25%	0.000000	18.000.000	2.016.000.000	16.000.000	8.000.000	0.000000	1.000.000	2.000.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	7.000.000	0.000000	69.500.000	0.000000	0.000000
50%	0.000000	69.000.000	2.016.000.000	28.000.000	16.000.000	1.000.000	2.000.000	2.000.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	9.000.000	0.000000	94.950.000	0.000000	0.000000
75%	1.000.000	161.000.000	2.017.000.000	38.000.000	23.000.000	2.000.000	3.000.000	2.000.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	152.000.000	0.000000	126.000.000	0.000000	1.000.000
max	1.000.000	737.000.000	2.017.000.000	53.000.000	31.000.000	19.000.000	50.000.000	55.000.000	10.000.000	10.000.000	1.000.000	26.000.000	72.000.000	18.000.000	535.000.000	391.000.000	5.400.000.000	8.000.000	5.000.000



Gambar 4 tampilan matriks *missing values*



Gambar 5 tampilan *missing values*



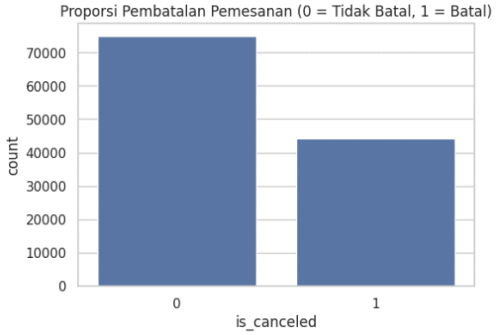
Gambar 6 tampilan hasil data *cleansing*

Dari analisis 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.

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 :

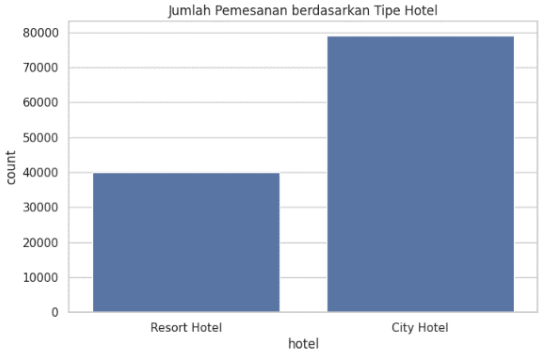
1. **Proporsi Pembatalan**



Gambar 7 tampilan presentase pembatalan

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

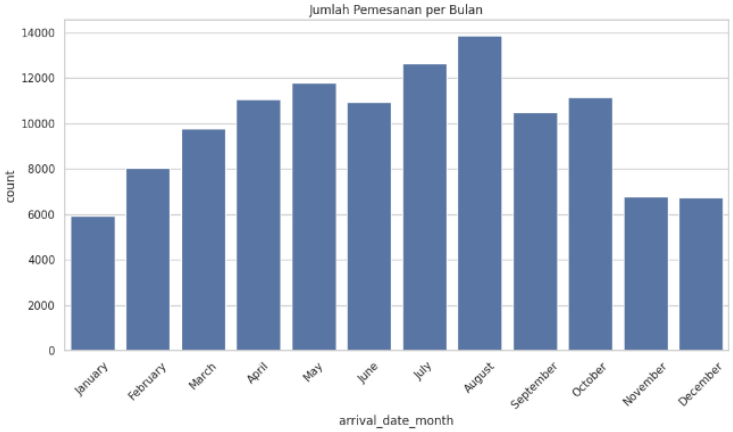
2. Tipe Hotel



Gambar 8 tampilan presentase tipe hotel

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

3. Pola Musiman



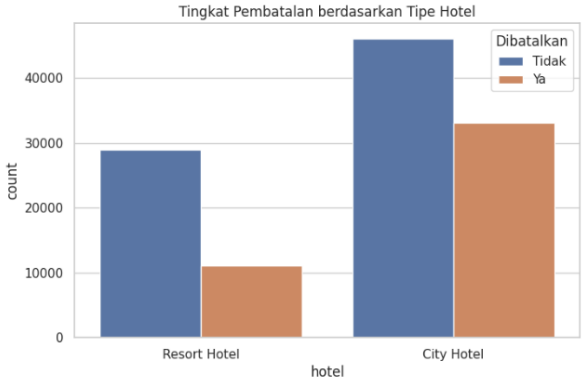
Gambar 9 tampilan kedatangan bulanan

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.

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.

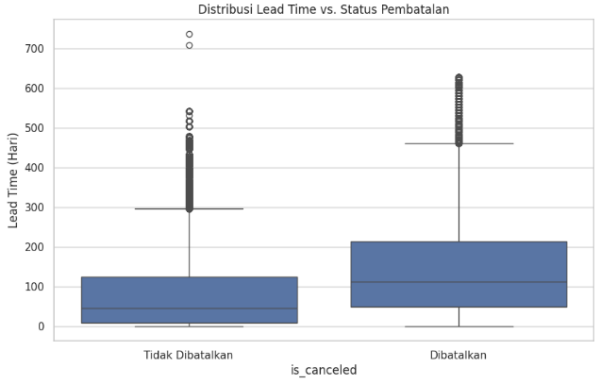
1. Tipe Hotel & Pembatalan



Gambar 10 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%).

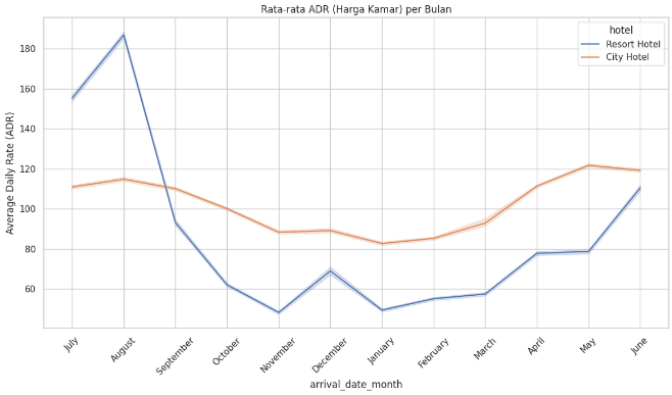
2. Waktu Tunggu (*Lead Time*) & Pembatalan



Gambar 11 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.

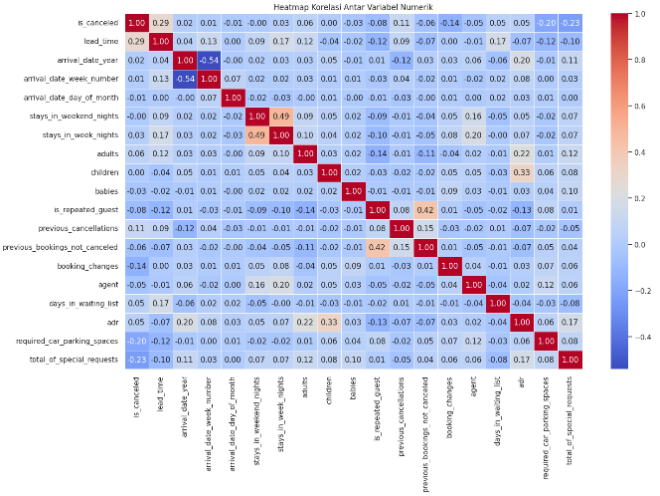
3. Tren Harga (ADR) Bulanan



Gambar 12 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.

4. Analisis Korelasi



Gambar 13 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.

PEMBAHASAN

Temuan bahwa 37,04% pemesanan dibatalkan menunjukkan bahwa pembatalan adalah masalah signifikan bagi industri perhotelan. Analisis bivariat menunjukkan bahwa tamu yang melakukan pemesanan jauh-jauh hari (*lead time* panjang) lebih cenderung untuk membatalkan dibandingkan tamu yang memesan mendekati tanggal kedatangan. Ini bisa menunjukkan adanya keraguan dalam rencana atau pemesanan yang bersifat spekulatif.

Hasil paling signifikan dari analisis korelasi adalah adanya hubungan negatif antara permintaan khusus dan pembatalan. Dimana tamu yang memiliki permintaan khusus, ex: *connecting room, high floor* jauh lebih kecil kemungkinan untuk melakukan pembatalan. Ini berarti bahwa tamu yang telah menghabiskan "investasi" waktu untuk menjelaskan kebutuhannya memiliki keinginan yang lebih besar untuk menginap.

KESIMPULAN

Analisis Data Eksploratif (EDA) pada dataset "Hotel Booking Demand" telah berhasil mengidentifikasi berbagai pola dan infomasi penting. Hasil utama mengungkapkan bahwa tingkat pembatalan secara keseluruhan sangat signifikan (37,04%), dengan persentase yang jauh lebih besar di *City* Hotel (41,7%) dibandingkan *Resort* Hotel.

Faktor prediktif paling signifikan untuk pembatalan adalah waktu tunggu/*lead time* yang panjang, dimana semakin besar jarak antara pemesanan dan *check-in*, semakin tinggi kemungkinan pembatalan. Sebaliknya, partisipasi tamu, yang tercermin dalam *total_of_special_requests*, secara signifikan menurunkan risiko pembatalan. Secara musiman, pemesanan dan tarif kamar mencapai puncaknya pada bulan-bulan musim panas, terutama pada bulan Agustus.

Hasil ini dapat dimanfaatkan oleh manajemen hotel dalam merancang strategi mitigasi, seperti menerapkan kebijakan deposit yang lebih ketat untuk pemesanan dengan waktu tunggu yang lama atau melakukan pendekatan proaktif kepada tamu yang tidak memiliki permintaan tertentu.

UCAPAN TERIMA KASIH

Saya ingin mengungkapkan rasa terima kasih yang tulus kepada orang-orang hebat yang telah mendukung saya menyelesaikan paper ini.

Terima kasih tak terhingga untuk Ayah dan Ibu tercinta, yang tidak pernah lelah mendoakan, mendukung, dan berusaha keras membiayai seluruh pendidikan saya. Tanpa upaya Ayah dan Ibu, saya tidak akan berada di titik ini.

Saya juga ingin mengucapkan terima kasih kepada Ibu Karina Aulia Sari. Bimbingan serta pengetahuan mengenai EDA (Analisis Data Eksploratif) yang Ibu berikan sangat mengesankan dan memperluas wawasan saya.

Dan untuk Eva Ristiyanti, *support system* terbaik saya. Terima kasih selalu ada dan menjadi sumber semangat dan inspirasi saya. Dukungan, pengertian, dan kesabarannya sangat berarti untuk memberikan semangat dalam menyelesaikan paper ini.

Saya menyadari bahwa tanpa bantuan dan bimbingan dari berbagai pihak, penelitian ini tidak akan dapat terselesaikan dengan baik.

REFERENSI

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.

Aggarwal, C. C., & Sathe, S. (2017). Which outlier detection algorithm should I use?. In Outlier Ensembles: An Introduction (pp. 207-274). Cham: Springer International Publishing.

Kimes, S. E. (1989). The basics of yield management. Cornell Hotel and Restaurant Administration Quarterly, 30(3), 14-19.

Ivanov, S., & Zhechev, V. (2012). Hotel revenue management in theory and practice. Tourism & Management Studies International Conference, 155-165.

Mostipak, J. (2020). Hotel Booking Demand Dataset EDA. Kaggle. Diambil dari https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

McKinney, W. (2010). Data structures for statistical computing in Python. scipy, 445(1), 51-56.

Waskom, M. L. (2021). Seaborn: statistical data visualization. Journal of open source software, 6(60), 3021.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in science & engineering, 9(03), 90-95.

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and. Techniques, Waltham: Morgan Kaufmann Publishers.

Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.".

Camizuli, E., & Carranza, E. J. (2018). Exploratory data analysis (EDA). The encyclopedia of archaeological sciences, 1-7.