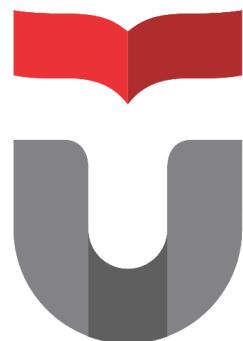


**‘TUGAS BESAR – GENAP 2024/2025**

**“BBK3BAB3 - Data Warehouse dan Business Intelligence”**

**Perancangan dan Implementasi Data Warehouse & Business Intelligence  
untuk Analisis Pemesanan Hotel**



**Telkom  
University**

**Disusun Oleh:**

Ahmad Fauzi 1202220263

Nerlis Fitria Nurani 1202223307

Rafie Safaraz Aribowo 1202223025

Ryannisa Syarifa Triandini 1202223163

Sarah Luki Raihani 1202223084

**PROGRAM STUDI SISTEM INFORMASI  
FAKULTAS REKAYASA INDUSTRI  
UNIVERSITAS TELKOM  
BANDUNG  
2025**

# DAFTAR ISI

<b>DAFTAR ISI.....</b>	<b>1</b>
<b>BAB I PENDAHULUAN.....</b>	<b>3</b>
1.1. Latar Belakang.....	3
1.2. Sasaran Strategis dan Indikator Kinerja.....	4
1.2.1. Objektif.....	4
1.2.1.1. Balanced Scorecard.....	4
1.2.3. KPI (Key Performance Indicator).....	4
<b>BAB II ANALISIS SUMBER DATA.....</b>	<b>9</b>
2.1. Sumber Data Utama.....	9
2.2. Exploratory Data Analysis (EDA).....	11
2.2.1. Struktur dan Tipe Data.....	11
2.2.2. Statistik Deskriptif.....	12
2.2.3. Analisis Missing Value.....	14
2.2.4. Analisis Unique Value.....	14
2.2.5. Analisis Korelasi.....	15
2.2.6. Analisis Outlier.....	17
2.2.7. Distribusi Pembatalan Reservasi Hotel.....	18
2.2.8. Tingkat Pembatalan Berdasarkan Jenis Hotel.....	19
2.2.9. Hubungan Lead Time dan Tingkat Pembatalan Hotel.....	19
2.2.10. Pola Musiman Pembatalan Berdasarkan Bulan Kedatangan.....	20
2.2.11. Pembatalan Berdasarkan Negara Asal Tamu.....	21
2.2.12. Distribusi ADR Berdasarkan Status Pembatalan.....	21
2.2.13. Hubungan Tipe Deposit dengan Tingkat Pembatalan.....	22
2.2.14. Dampak Riwayat Booking Terhadap Perilaku Pembatalan.....	23
2.2.15. Market Segment dalam Reservasi Hotel.....	24
<b>BAB III PERANCANGAN STAR SCHEMA.....</b>	<b>25</b>
3.1. Fact Table (fact_booking).....	26
3.2. Dimension Table.....	28
3.3. Measure dalam Analisis Data Warehouse Hotel Booking.....	32
<b>BAB IV IMPLEMENTASI STAR SCHEMA DALAM RDBMS.....</b>	<b>34</b>
4.1. Pembuatan Database.....	34
4.2. Pembuatan Dimension Table.....	34
4.3. Pembuatan Dimension Table.....	36
4.4. Desain Struktur Database DDL (RDBMS).....	40
<b>BAB V IMPLEMENTASI PROSES ETL.....</b>	<b>41</b>
5.1. Fact Table.....	41
5.2. Dimension Table.....	43
<b>BAB VI IMPLEMENTASI DATA MINING.....</b>	<b>66</b>
6.1. Klasifikasi (CatBoost Classifier).....	66

6.1.1. Data Cleaning & Preparation.....	66
6.1.2. Modelling (CatBoost Classifier).....	72
6.1.3. Hyperparameter Tuning dengan Optuna.....	73
6.2. Klusterisasi (K-Means).....	77
6.2.1 Pre-Modelling.....	77
6.3.2 Build K-Means Model.....	80
6.3.3 Export K-Means Model.....	82
<b>BAB VII PERANCANGAN DASHBOARD KPI.....</b>	<b>83</b>
7.1. Implementasi Dashboard Visualisasi KPI.....	83
7.2. Analisis & Evaluasi KPI.....	87
7.2.1 Average Length of Stay (LOS).....	87
7.2.2 Special Request Fulfillment Rate.....	88
7.2.3 Year-over-Year Revenue Growth.....	90
7.2.4 Repeat Guest Rate.....	91
7.2.5 Suite Booking Rate.....	92
7.2.6 Cancellation Rate.....	93
7.2.7 Cancellation Rate by Deposit Type.....	95
7.2.8 Country Cancellation Rate.....	97
7.2.9 Market Segment Conversion Rate.....	98
7.2.10 Lead Time Trend by Arrival Date Year.....	101
<b>BAB VIII KESIMPULAN DAN SARAN.....</b>	<b>103</b>
8.1. Kesimpulan.....	103
8.2. Saran.....	103
<b>PENGERJAAN TUGAS ANGGOTA KELOMPOK.....</b>	<b>104</b>
<b>LAMPIRAN.....</b>	<b>107</b>

# BAB I PENDAHULUAN

## 1.1. Latar Belakang

Dalam era digital yang terus berkembang, sektor perhotelan menghadapi tantangan signifikan dalam mengelola data operasional untuk mengoptimalkan keputusan bisnis. Hotel, sebagai salah satu komponen utama industri pariwisata, menghasilkan volume data yang besar dari berbagai aktivitas seperti pemesanan kamar, layanan pelanggan, dan manajemen operasional. Namun, data ini seringkali tersebar di berbagai sistem dan belum terintegrasi secara optimal, sehingga menghambat kemampuan manajemen untuk mendapatkan *insights* yang komprehensif.

Implementasi *Data Warehouse* dan *Business Intelligence* (DWBI) menjadi solusi strategis untuk mengatasi permasalahan tersebut. Dengan mengintegrasikan data dari berbagai sumber ke dalam satu repository terpusat, hotel dapat melakukan analisis yang lebih mendalam dan menghasilkan keputusan berbasis data yang lebih akurat. Dataset "*Hotel Booking Demand*" yang digunakan dalam proyek ini menyediakan informasi komprehensif tentang pemesanan hotel, termasuk aspek-aspek seperti waktu pemesanan, durasi menginap, jumlah tamu, dan berbagai parameter operasional lainnya.

*Star schema* yang dirancang dalam proyek ini bertujuan untuk memfasilitasi analisis *multidimensional* terhadap data pemesanan hotel. Struktur ini terdiri dari tabel fakta (fact\_booking) yang berhubungan dengan beberapa tabel dimensi seperti dim\_date, dim\_guest, dim\_hotel, dim\_room, dan dimensi lainnya. Pendekatan ini memungkinkan analisis kompleks dari berbagai perspektif bisnis, seperti tren pemesanan berdasarkan waktu, segmentasi pelanggan, atau performa hotel berdasarkan lokasi.

Melalui proses *Extract, Transform, Load* (ETL), data mentah dari sistem operasional dikonversi menjadi format yang sesuai untuk analisis. Proses ini melibatkan pembersihan data, transformasi struktur, dan pengayaan dengan *metadata* tambahan untuk meningkatkan nilai analitisnya. Hasil akhir dari proses ETL adalah *data warehouse* yang terstruktur dengan baik dan siap untuk eksplorasi melalui berbagai *tools Business Intelligence*.

Selain analisis deskriptif, proyek ini juga akan menerapkan teknik *Data Mining* untuk mengungkap pola tersembunyi dan hubungan yang tidak terlihat secara langsung dari data. Pendekatan ini memungkinkan hotel untuk mendapatkan *insights* prediktif seperti faktor-faktor yang mempengaruhi pembatalan pemesanan, perilaku segmen pelanggan tertentu, dan lain-lainnya.

Dengan mengadopsi pendekatan *Balanced Scorecard*, proyek ini akan menerjemahkan *insights* yang diperoleh menjadi *Key Performance Indicators* (KPI) yang terukur dan relevan dengan tujuan bisnis hotel. KPI ini akan menjadi dasar untuk pembuatan *dashboard* interaktif yang menyajikan visualisasi data secara komprehensif, memungkinkan

*stakeholders* untuk memantau performa bisnis secara *real-time* dan membuat keputusan yang tepat waktu.

Implementasi DWBI dalam konteks industri perhotelan tidak hanya meningkatkan efisiensi operasional tetapi juga memberikan keunggulan kompetitif melalui pengambilan keputusan yang lebih cerdas dan responsif terhadap perubahan kondisi pasar. Dengan infrastruktur analitik yang kuat, hotel dapat meningkatkan kualitas layanan, mengoptimalkan strategi penetapan harga, dan pada akhirnya meningkatkan profitabilitas secara keseluruhan.

## 1.2. Sasaran Strategis dan Indikator Kinerja

### 1.2.1. Objektif

Dashboard ini dirancang untuk memberikan gambaran komprehensif mengenai kinerja operasional dan finansial hotel secara real time, sehingga manajemen dapat dengan cepat mengidentifikasi tren, mengukur efektivitas strategi, dan mengambil keputusan berbasis data. Dengan menghadirkan visualisasi indikator utama mulai dari durasi menginap hingga tingkat pembatalan dan segmen pasar dashboard membantu memastikan bahwa target pendapatan, efisiensi operasional, dan loyalitas tamu dapat dicapai secara optimal tanpa perlu mengurai detail teknis setiap metrik secara manual.

#### 1.2.1.1. Balanced Scorecard

Dalam framework Balanced Scorecard, dashboard ini mengintegrasikan empat perspektif utama:

- Keuangan (Financial): Mengukur pertumbuhan pendapatan (YoY Revenue Growth), margin kamar premium (Suite Room Booking Rate), dan efisiensi biaya operasional (Average LOS).
- Pelanggan (Customer): Memantau kepuasan dan loyalitas melalui Repeat Guest Rate serta konversi reservasi khusus (Special Requests Fulfillment Rate).
- Proses Internal (Internal Process): Menilai efektivitas kebijakan deposit dan lead time pemesanan (Cancellation Rate by Deposit Type, Lead Time Trend) untuk mengurangi pembatalan dan meningkatkan utilisasi kamar.
- Pembelajaran & Pertumbuhan (Learning & Growth): Melacak adopsi praktik terbaik dan inisiatif inovasi dengan melihat tren permintaan khusus dan segmentasi pasar (Market Segment Conversion Rate), sebagai tolok ukur kesiapan tim dalam menyesuaikan layanan dan strategi revenue management.

Setiap perspektif saling melengkapi, memastikan bahwa sasaran finansial tidak dicapai dengan mengorbankan kualitas layanan, dan sebaliknya, upaya peningkatan operasional dan pengembangan SDM dapat dikaitkan langsung dengan hasil bisnis.

#### 1.2.3. KPI (Key Performance Indicator)

No	Goals	KPI	Data	Perhitungan	Chart
----	-------	-----	------	-------------	-------

1.	<b>Average Length of Stay (LOS)</b>  Meningkatkan durasi rata-rata menginap tamu untuk mengoptimalkan pendapatan hotel dan mengurangi biaya operasional per malam	Average Length of Stay (LOS) mencapai minimal 2 malam per reservasi untuk meningkatkan efisiensi operasional dan pendapatan total hotel	fact_booking: stays_in_weekend_nights, stays_in_week_nights, is_canceled	<b>Actual:</b> SUM(CASE WHEN is_canceled = 0 THEN stays_in_weekend_nights + stays_in_week_nights ELSE 0 END) / SUM(CASE WHEN is_canceled = 0 THEN 1 ELSE 0 END)  <b>Skala Warna Target:</b> - Merah: 0.5 malam - Kuning: 1 malam - Hijau: 2 malam	Text Chart
2.	<b>Special Request Fulfillment Rate</b>  Meningkatkan tingkat pemenuhan permintaan khusus pelanggan untuk meningkatkan kepuasan pelanggan.	Special Request Fulfillment Rate mencapai minimal 50% untuk memastikan mayoritas permintaan khusus pelanggan terpenuhi dengan baik	fact_booking: total_of_special_requests, is_canceled	<b>Actual:</b> (SUM(CASE WHEN total_of_special_requests > 0 AND is_canceled = 0 THEN 1 ELSE 0 END) / SUM(CASE WHEN total_of_special_requests > 0 THEN 1 ELSE 0 END))  <b>Skala Warna Target:</b> - Merah: 10% - Kuning: 30% - Hijau: 50%	Text Chart
3.	<b>Year-over-Year Revenue Growth</b>  Memastikan pertumbuhan pendapatan hotel secara berkelanjutan dan melebihi rata-rata industri untuk memperkuat	Year-over-Year Revenue Growth mencapai minimal 3% untuk mengungguli rata-rata pertumbuhan industri perhotelan dan memastikan keberlanjutan bisnis jangka panjang	fact_booking: revenue  dim_date: year	<b>Actual:</b> ((Total Revenue Tahun 2016 - Total Revenue Tahun 2015) / Total Revenue Tahun 2015) × 100%  <b>Target:</b> ≥3% pertumbuhan pendapatan YoY	Text Chart

	posisi kompetitif di pasar				
4.	<b>Repeat Guest Rate</b>  Meningkatkan loyalitas pelanggan dan profitabilitas jangka panjang melalui peningkatan jumlah tamu yang kembali menginap	Repeat Guest Rate mencapai minimal 15% untuk membangun basis pelanggan loyal yang menghasilkan pendapatan berkelanjutan dan mengurangi biaya akuisisi	fact_booking: is_repeated_guest	<b>Actual:</b> (SUM(CASE WHEN is_repeated_guest = 1 THEN 1 ELSE 0 END) / COUNT(*))  <b>Skala Warna Target:</b> - Merah: 5% - Kuning: 10% - Hijau: 15%	Bullet chart
5.	<b>Suite Booking Rate</b>  Meningkatkan proporsi pemesanan kamar hotel tipe Suite untuk memaksimalkan pendapatan per kamar tersedia dan meningkatkan profitabilitas hotel secara keseluruhan	Suite Booking Rate mencapai minimal 5% dari total pemesanan untuk mengoptimalkan revenue mix dan meningkatkan Average Daily Rate (ADR) hotel	dim_room : reserved_room_type  fact_booking : is_canceled	<b>Actual:</b> (SUM(CASE WHEN reserved_room_category = 'G' THEN 1 ELSE 0 END) / SUM(CASE WHEN is_canceled = 0 THEN 1 ELSE 0 END))  <b>Skala Warna Target:</b> - Merah: 1% - Kuning: 2,5% - Hijau: 5%	Bullet Chart
6.	<b>Cancellation Rate</b>  Meminimalkan tingkat pembatalan reservasi untuk mengoptimalkan	Cancellation Rate maksimal 15% dari total reservasi untuk menjaga stabilitas	fact_booking: is_canceled	<b>Actual:</b> (SUM(CASE WHEN is_canceled = 1 THEN 1 ELSE 0 END) / COUNT(is_canceled))  <b>Range Target:</b>	Gauge Chart

	perencanaan operasional dan memaksimalkan pendapatan hotel	pendapatan dan efisiensi perencanaan operasional hotel		- Merah: 0-25% - Kuning: 25-50% - Hijau: 50-100%	
7.	<p><b>Cancellation Rate by Deposit Type</b></p> <p>Mengoptimalkan kebijakan deposit hotel untuk meminimalisir pembatalan reservasi dan memaksimalkan pendapatan terjamin</p>	<p>Cancellation Rate by Deposit Type dengan target pembatalan maksimal:</p> <p>No Deposit: <math>\leq 25\%</math></p> <p>Refundable: <math>\leq 15\%</math></p> <p>Non-Refundable: <math>\leq 5\%</math></p> <p>untuk menyeimbangkan antara fleksibilitas bagi tamu dan keamanan pendapatan hotel</p>	<p>dim_deposit_type : deposit_type, fact_table : is_canceled</p>	<p><b>Actual:</b>  <math>(\text{SUM}(\text{CASE WHEN is\_canceled = 1 THEN 1 ELSE 0 END}) / \text{COUNT(is\_canceled)})</math></p> <p>Namun pada properti gauge chartnya difilter berdasarkan masing-masing tipe deposit.</p> <p><b>Range Target:</b></p> <p>No Deposit</p> <ul style="list-style-type: none"> <li>- Merah: 0-25%</li> <li>- Kuning: 25-50%</li> <li>- Hijau: 50-100%</li> </ul> <p>Refundable</p> <ul style="list-style-type: none"> <li>- Merah: 0-15%</li> <li>- Kuning: 15-50%</li> <li>- Hijau: 50-100%</li> </ul> <p>Non-Refundable</p> <ul style="list-style-type: none"> <li>- Merah: 0-5%</li> <li>- Kuning: 5-50%</li> <li>- Hijau: 50-100%</li> </ul>	Gauge chart
8.	<p><b>Country Cancellation Rate</b></p> <p>Mengidentifikasi dan mengoptimalkan tingkat pembatalan berdasarkan negara asal tamu untuk</p>	<p>Country Cancellation Rate dengan target maksimal 25% pembatalan untuk setiap negara utama, memungkinkan penargetan strategi</p>	<p>dim_country : country_name fact_booking : is_canceled</p>	<p><b>Actual:</b>          Dimensi Geografis menggunakan country_name. Metrik menggunakan Cancellation Rate dengan rumus perhitungan yang sama dengan KPI 6 dan 7.</p> <p><b>Range Warna Target:</b></p> <ul style="list-style-type: none"> <li>- Nilai Warna Min: Hijau</li> </ul>	Map Chart

	mengembangkan strategi pemasaran dan kebijakan reservasi yang terpersonalisasi	retensi pemesanan yang efektif berdasarkan pasar geografis		- Nilai Warna Tengah: Kuning - Nilai Warna Max: Merah	
9.	<b>Number of Customers by Room Rate</b>  Mengoptimalkan strategi penetapan harga dengan mengidentifikasi preferensi pelanggan berdasarkan kategori tarif kamar.	Minimal 50% pelanggan berada pada kategori Low Rate untuk memaksimal kan okupansi dan daya tarik pasar.	dim_market_segment: market_segment  fact_booking: is_canceled	<b>Actual:</b> Hasil dari data mining pada metode Clustering menggunakan K-Means. Dibagi kedalam 3 K sesuai dengan optimal elbow point.  <b>Target:</b> $(\text{Jumlah pelanggan di kategori Low Rate}) / \text{Total pelanggan} = 78.187 / (78.187 + 25.190 + 13.136) \approx 67\%$	Bar Chart
10.	<b>Lead Time Trend by Arrival Date Year</b>  Meningkatkan rata-rata lead time pemesanan untuk mengoptimalkan perencanaan operasional dan meningkatkan stabilitas pendapatan hotel	Average Lead Time minimal 100 hari, dengan target pertumbuhan tahunan 10%, untuk memberikan visibilitas jangka panjang bagi manajemen inventaris dan strategi harga dinamis	fact_booking: lead_time  dim_date: year	<b>Actual:</b> $\text{SUM(lead\_time)} / \text{COUNT(*)}$  dengan dimensi year  <b>Target:</b> $\geq 100$ hari (Lead time optimal)	Line Chart

## BAB II ANALISIS SUMBER DATA

### 2.1. Sumber Data Utama

Sumber data utama yang digunakan dalam proyek ini berasal dari dataset "Hotel Booking Demand" yang diakses melalui platform Kaggle. Dataset ini berisi informasi komprehensif tentang pemesanan hotel untuk dua jenis properti - City Hotel dan Resort Hotel, mencakup berbagai atribut penting seperti waktu pemesanan, durasi menginap, karakteristik tamu, metode distribusi pemesanan, dan banyak faktor lain yang relevan dengan analisis bisnis perhotelan.

Tautan Dataset
<a href="https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand?select=hotel_bookings.csv">https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand?select=hotel_bookings.csv</a>

Dataset ini berasal dari penelitian akademik yang dipublikasikan dalam artikel ilmiah "Hotel Booking Demand Datasets" yang ditulis oleh Nuno Antonio, Ana Almeida, dan Luis Nunes dalam jurnal Data in Brief, Volume 22, Februari 2019. Data kemudian diunduh dan dibersihkan oleh Thomas Mock dan Antoine Bichat untuk proyek #TidyTuesday pada minggu 11 Februari 2020. Semua informasi yang dapat mengidentifikasi individu telah dihapus dari dataset ini untuk menjaga privasi. Dataset ini memiliki 119.390 entri dengan 32 kolom/fitur yang mencakup informasi tentang:

Nama Kolom	Tipe Data	Deskripsi
hotel	Kategorikal	Jenis hotel tempat pemesanan dilakukan (City Hotel atau Resort Hotel)
is_canceled	Biner (0/1)	Indikator apakah pemesanan dibatalkan (1) atau tidak (0)
lead_time	Numerik	Jumlah hari antara tanggal pemesanan dan tanggal kedatangan tamu
arrival_date_year	Numerik	Tahun kedatangan tamu
arrival_date_month	Kategorikal	Bulan kedatangan tamu (Januari hingga Desember)
arrival_date_week_number	Numerik	Nomor minggu dalam tahun untuk tanggal kedatangan tamu

arrival_date_day_of_month	Numerik	Hari dalam bulan untuk tanggal kedatangan tamu
stays_in_weekend_nights	Numerik	Jumlah malam menginap pada akhir pekan (Sabtu-Minggu) yang dipesan
stays_in_week_nights	Numerik	Jumlah malam menginap pada hari kerja (Senin-Jumat) yang dipesan
adults	Numerik	Jumlah tamu dewasa dalam pemesanan
children	Numerik	Jumlah anak-anak dalam pemesanan
babies	Numerik	Jumlah bayi dalam pemesanan
meal	Kategorikal	Jenis paket makanan yang dipesan (BB: Bed & Breakfast, HB: Half Board, FB: Full Board, SC: Self Catering)
country	Kategorikal	Negara asal tamu (kode ISO 3)
market_segment	Kategorikal	Segmen pasar yang digunakan untuk pemesanan (Direct, Corporate, Online TA, Offline TA/TO, Groups, dll)
distribution_channel	Kategorikal	Saluran distribusi yang digunakan untuk pemesanan (Direct, Corporate, TA/TO, dll)
is_repeated_guest	Biner (0/1)	Indikator apakah tamu pernah menginap sebelumnya (1) atau tidak (0)
previous_cancellations	Numerik	Jumlah pemesanan sebelumnya yang dibatalkan oleh tamu
previous_bookings_not_canceled	Numerik	Jumlah pemesanan sebelumnya yang tidak dibatalkan oleh tamu
reserved_room_type	Kategorikal	Kode jenis kamar yang dipesan
assigned_room_type	Kategorikal	Kode jenis kamar yang sebenarnya diberikan kepada tamu

booking_changes	Numerik	Jumlah perubahan/permintaan yang dilakukan pada pemesanan
deposit_type	Kategorikal	Jenis deposit yang dibayarkan (No Deposit, Non Refund, Refundable)
agent	Numerik	ID agen travel yang melakukan pemesanan, jika ada
company	Numerik	ID perusahaan yang membuat pemesanan, jika ada
days_in_waiting_list	Numerik	Jumlah hari pemesanan berada dalam daftar tunggu sebelum dikonfirmasi
customer_type	Kategorikal	Jenis pemesanan (Transient, Contract, Group, Transient-Party)
adr	Numerik	Average Daily Rate - rata-rata harga per hari dari pemesanan
required_car_parking_spaces	Numerik	Jumlah tempat parkir mobil yang diminta oleh tamu
total_of_special_requests	Numerik	Jumlah permintaan khusus yang diajukan oleh tamu
reservation_status	Kategorikal	Status terakhir reservasi (Check-Out, Canceled, No-Show)
reservation_status_date	Tanggal	Tanggal saat status pemesanan terakhir diperbarui

## 2.2. Exploratory Data Analysis (EDA)

### 2.2.1. Struktur dan Tipe Data

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

Dataset Hotel Booking Demand memiliki 119.390 entri dengan 32 kolom yang terdiri dari berbagai tipe data. Sebagian besar kolom (16 kolom) bertipe data integer (int64) yang mewakili nilai numerik diskrit seperti jumlah malam menginap, jumlah tamu, dan permintaan khusus. Terdapat 12 kolom dengan tipe data objek (object) yang umumnya mewakili data kategorikal seperti jenis hotel, negara asal, dan status reservasi. Sisanya adalah 4 kolom dengan tipe data float (float64) yang digunakan untuk nilai desimal seperti ADR (Average Daily Rate) atau untuk kolom yang memiliki nilai null seperti children, agent, dan company.

Keseluruhan dataset memiliki kelengkapan data yang baik, dengan mayoritas kolom memiliki 119.390 nilai non-null. Namun terdapat beberapa kolom dengan nilai yang hilang, seperti children (119.386 non-null), country (118.902 non-null), agent (103.050 non-null), dan company (6.797 non-null). Kolom company memiliki jumlah data yang sangat sedikit, hanya sekitar 5,7% dari total data, yang mengindikasikan bahwa sebagian besar pemesanan tidak terkait dengan akun perusahaan. Kolom agent juga memiliki cukup banyak nilai yang hilang, yang menunjukkan bahwa tidak semua pemesanan dilakukan melalui agen travel. Pola ketidaklengkapan data ini memberikan wawasan awal tentang karakteristik pemesanan hotel, di mana pemesanan langsung dari pelanggan individual tampaknya lebih umum dibandingkan pemesanan korporat atau melalui agen.

## 2.2.2. Statistik Deskriptif

Dataset ini berisi 119.390 data pemesanan hotel dengan berbagai fitur numerik dan kategorikal yang berkaitan dengan informasi reservasi, tamu, dan kondisi pemesanan. Dari hasil eksplorasi data awal (EDA), ditemukan sejumlah insight penting.

### a. Fitur Numerik

	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
count	119390.000000	119390.000000	119390.000000	119390.000000	27.165173	15.798241	0.927599	2.500302	1.856403	0.103880	0.007949	0.031912
mean	0.370416	104.011416	2016.156554									0.087118
std	0.482918	106.863097	0.707476	13.605138	8.780829	0.998613	1.908286	0.579261	0.398561	0.097436	0.175767	0.844236
min	0.000000	0.000000	2015.000000		1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	0.000000	1.000000	2.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	1.000000	2.000000	2.000000	0.000000	0.000000	0.000000	0.000000
75%	1.000000	160.000000	2017.000000	38.000000	23.000000	2.000000	3.000000	2.000000	0.000000	0.000000	0.000000	0.000000
max	1.000000	737.000000	2017.000000	53.000000	31.000000	19.000000	50.000000	55.000000	10.000000	10.000000	1.000000	26.000000

previous_bookings_not_canceled	booking_changes	agent	company	days_in_waiting_list	adr	required_car_parking_spaces	total_of_special_requests
119390.000000	119390.000000	103050.000000	6797.000000	119390.000000	119390.000000	119390.000000	119390.000000
0.137097	0.221124	86.693382	189.266735	2.321149	101.831122	0.062518	0.571363
1.497437	0.652306	110.774548	131.655015	17.594721	50.535790	0.245291	0.792798
0.000000	0.000000	1.000000	6.000000	0.000000	-6.380000	0.000000	0.000000
0.000000	0.000000	9.000000	62.000000	0.000000	69.290000	0.000000	0.000000
0.000000	0.000000	14.000000	179.000000	0.000000	94.575000	0.000000	0.000000
0.000000	0.000000	229.000000	270.000000	0.000000	126.000000	0.000000	1.000000
72.000000	21.000000	535.000000	543.000000	391.000000	5400.000000	8.000000	5.000000

- Pembatalan Reservasi (is\_canceled) menunjukkan bahwa sekitar 37% pemesanan dibatalkan (mean = 0.37), sedangkan sisanya dilanjutkan.
- lead\_time memiliki rata-rata lebih dari 100 hari, menunjukkan bahwa tamu sering memesan jauh hari sebelumnya. Nilai maksimal mencapai 737 hari.
- Durasi Menginap terbagi menjadi stays\_in\_weekend\_nights dan stays\_in\_week\_nights, dengan rata-rata masing-masing sekitar 0.9 dan 2.5 malam, namun beberapa tamu tinggal hingga 19 malam di akhir pekan dan 50 malam di hari kerja.
- Fitur adults, children, dan babies mengindikasikan bahwa sebagian besar pemesanan dilakukan oleh dua orang dewasa tanpa anak-anak atau bayi, meskipun ada nilai ekstrim seperti 10 anak dan 10 bayi dalam satu pemesanan.
- is\_repeated\_guest memiliki nilai rata-rata yang sangat rendah (0.03), menunjukkan sebagian besar tamu adalah pelanggan baru.
- previous\_cancellations dan previous\_bookings\_not\_canceled menunjukkan bahwa sebagian besar tamu belum pernah membatalkan atau memiliki pemesanan sebelumnya.
- booking\_changes sangat kecil secara rata-rata, menandakan perubahan pemesanan jarang dilakukan.
- adr (Average Daily Rate) berkisar dari -6.38 hingga 5400, dengan rata-rata sekitar 101. Nilai negatif atau sangat tinggi menunjukkan adanya outlier atau anomali.
- required\_car\_parking\_spaces juga mayoritas bernilai nol, menunjukkan banyak tamu tidak membutuhkan parkir.
- total\_of\_special\_requests menunjukkan bahwa lebih dari setengah tamu mengajukan setidaknya satu permintaan khusus.

## b. Fitur Kategorikal

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reserved_room_type	assigned_room_type	deposit_type	customer_type	reservation_status	reservation_status_date
count	119390	119390	119390	118902	119390	119390	119390	119390	119390	119390	119390	119390
unique	2	12	5	177	8	5	10	12	3	4	3	926
top	City Hotel	August	BB	PRT	Online TA	TA/TO	A	A	No Deposit	Transient	Check-Out	2015-10-21
freq	79330	13877	92310	48590	56477	97870	85994	74053	104641	89613	75166	1461

- hotel terdiri dari dua jenis, dengan City Hotel menjadi pilihan dominan (79.330 data).
- arrival\_date\_month menunjukkan Agustus sebagai bulan kedatangan paling populer.
- meal paling umum adalah BB (Bed & Breakfast).
- country terbanyak adalah PRT (Portugal), diikuti oleh negara lain dengan total 177 negara unik.

- market\_segment paling sering adalah Online TA (Travel Agent), sedangkan distribution\_channel paling umum adalah TA/TO (Travel Agent/Tour Operator).
- reserved\_room\_type dan assigned\_room\_type paling banyak adalah tipe A, meskipun terdapat 10 jenis kamar yang berbeda.
- deposit\_type paling umum adalah No Deposit, menandakan tamu tidak diminta untuk memberikan deposit sebelumnya.
- customer\_type paling banyak adalah Transient, yakni tamu biasa tanpa kontrak perusahaan atau grup.
- reservation\_status menunjukkan mayoritas pemesanan berstatus Check-Out, berarti tamu telah menginap dan menyelesaikan masa menginapnya.

### 2.2.3. Analisis Missing Value

Tabel Informasi Missing Value

Fitur	Jumlah Missing	Persentase (%)
0 company	112593	94.310000
1 agent	16340	13.690000
2 country	488	0.410000
22 children	4	0.000000

Analisis missing value pada dataset Hotel Booking Demand menunjukkan pola ketidaklengkapan data yang bervariasi dan memberikan wawasan tentang karakteristik pemesanan hotel. Kolom "company" memiliki persentase missing value tertinggi (94,31% atau 112.593 entri), mengindikasikan bahwa mayoritas pemesanan dilakukan oleh individu bukan perusahaan. Kolom "agent" menunjukkan 13,69% (16.340) nilai yang hilang, mencerminkan proporsi pemesanan yang dilakukan tanpa perantara agen travel. Sementara kolom "country" memiliki missing value yang minimal (0,41% atau 488 entri), dan kolom "children" hampir lengkap dengan hanya 4 entri yang hilang (kurang dari 0,01%). Pola ini mengindikasikan bahwa dataset memiliki kelengkapan informasi demografis tamu yang baik, meskipun informasi afiliasi bisnis relatif terbatas, yang konsisten dengan karakteristik umum industri perhotelan rekreasional.

### 2.2.4. Analisis Unique Value

Tabel Informasi Unique Value per Kolom di df\_categorical

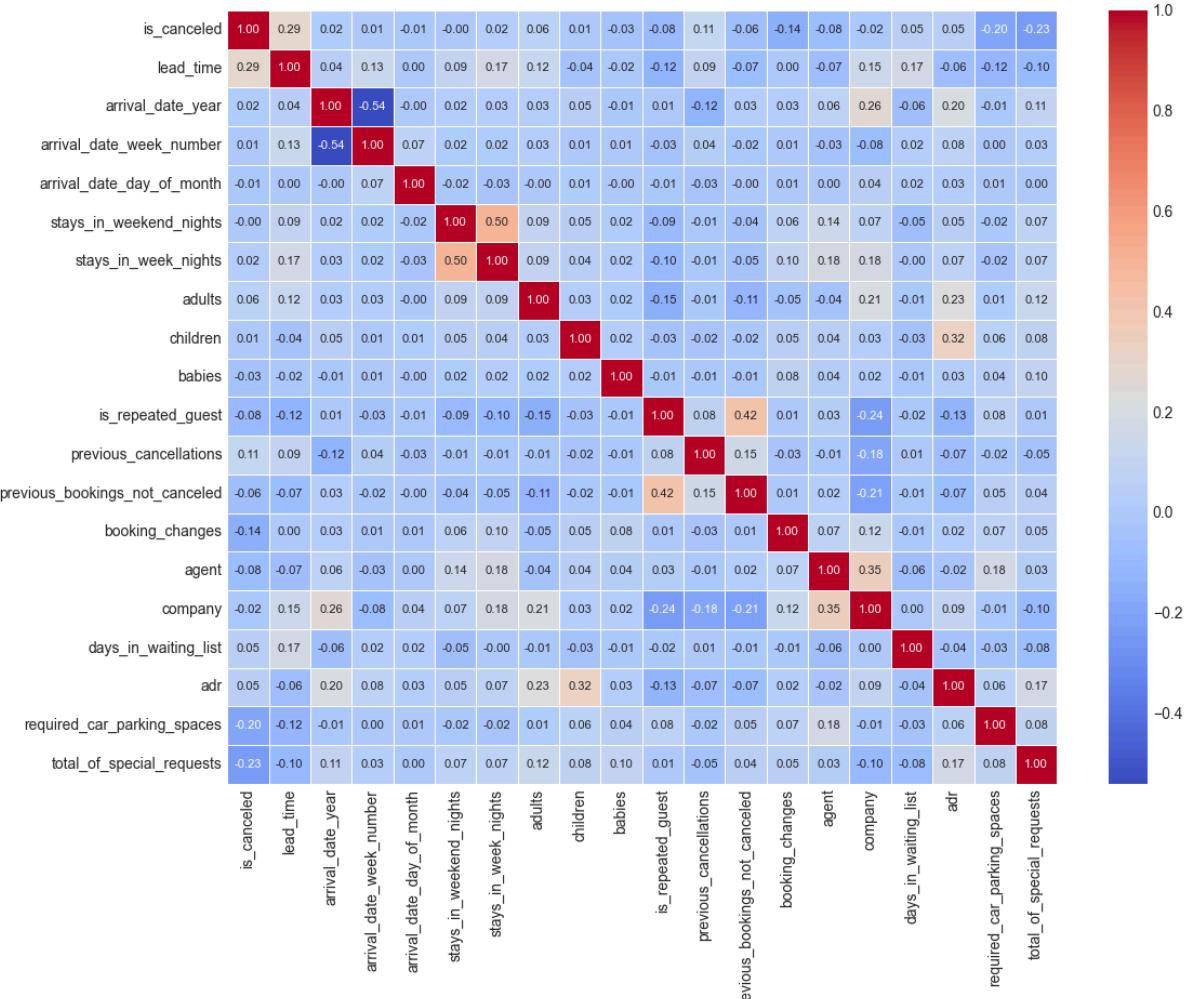
Fitur	Jumlah Unique	Persentase (%)	Tipe Data	Contoh Nilai Unik
0 reservation_status_date	926	0.780000	object	['2015-07-01' '2015-07-02' '2015-07-03' '2015-05-06' '2015-04-22' '2015-06-23' '2015-07-05' '2015-07-06' '2015-07-07' '2015-07-08']
1 country	177	0.150000	object	['PRY' 'GBR' 'USA' 'ESP' 'IRL' 'FRA' 'ROU' 'NOR' 'OMN' 'ARG']
2 arrival_date_month	12	0.010000	object	['July' 'August' 'September' 'October' 'November' 'December' 'January' 'February' 'March' 'April']
3 market_segment	8	0.010000	object	['Direct' 'Corporate' 'Online TA' 'Offline TA/TO' 'Complementary' 'Groups' 'Undefined' 'Aviation']
4 reserved_room_type	10	0.010000	object	['C' 'A' 'D' 'E' 'G' 'F' 'H' 'L' 'P' 'B']
5 assigned_room_type	12	0.010000	object	['C' 'A' 'D' 'E' 'G' 'F' 'I' 'B' 'H' 'P']
6 hotel	2	0.000000	object	['Resort Hotel' 'City Hotel']
7 meal	5	0.000000	object	['BB' 'FB' 'HB' 'SC' 'Undefined']
8 distribution_channel	5	0.000000	object	['Direct' 'Corporate' 'TA/TO' 'Undefined' 'GDS']
9 deposit_type	3	0.000000	object	['No Deposit' 'Refundable' 'Non Refund']
10 customer_type	4	0.000000	object	['Transient' 'Contract' 'Transient-Party' 'Group']
11 reservation_status	3	0.000000	object	['Check-Out' 'Canceled' 'No-Show']

Analisis nilai unik pada dataset Hotel Booking Demand mengungkapkan beberapa temuan penting untuk proses preprocessing data. Kolom "reservation\_status\_date" memiliki

jumlah nilai unik tertinggi (926) dan masih bertipe data object, bukan datetime, yang mengindikasikan perlunya konversi tipe data untuk memfasilitasi analisis temporal. Kolom "country" memiliki 177 nilai unik yang menunjukkan keragaman geografis pelanggan hotel. Beberapa kolom kategorikal seperti "meal", "market\_segment", dan "distribution\_channel" mengandung nilai 'Undefined' yang perlu diproses, dengan pendekatan terbaik adalah imputasi menggunakan modus (nilai yang paling sering muncul) untuk mempertahankan distribusi data. Terdapat perbedaan antara "reserved\_room\_type" (10 nilai unik) dan "assigned\_room\_type" (12 nilai unik) yang mengindikasikan adanya perubahan kamar pada saat check-in. Dataset ini juga mencakup dua tipe hotel ('Resort Hotel' dan 'City Hotel') serta empat tipe pelanggan ('Transient', 'Contract', 'Transient-Party', 'Group') yang dapat digunakan untuk segmentasi analisis. Kolom-kolom dengan cardinalitas rendah seperti "deposit\_type" (3 nilai) dan "reservation\_status" (3 nilai) menunjukkan data kategorikal yang terstruktur dengan baik dan siap untuk digunakan dalam pemodelan prediktif tanpa memerlukan transformasi lebih lanjut.

## 2.2.5. Analisis Korelasi

Correlation Heatmap - Hotel Booking Features



Berdasarkan heatmap korelasi pada gambar diatas, beberapa temuan penting terlihat dalam pola hubungan antar fitur dalam dataset Hotel Booking Demand:

Korelasi terkuat terlihat antara "stays\_in\_weekend\_nights" dan "stays\_in\_week\_nights" dengan nilai 0,50, yang menunjukkan hubungan positif moderat. Ini mengindikasikan bahwa tamu yang memesan untuk menginap di akhir pekan juga cenderung memesan untuk menginap di hari kerja, mencerminkan pola pemesanan untuk durasi menginap yang lebih panjang. Temuan ini sejalan dengan perilaku wisatawan yang sering merencanakan perjalanan melampaui akhir pekan saja.

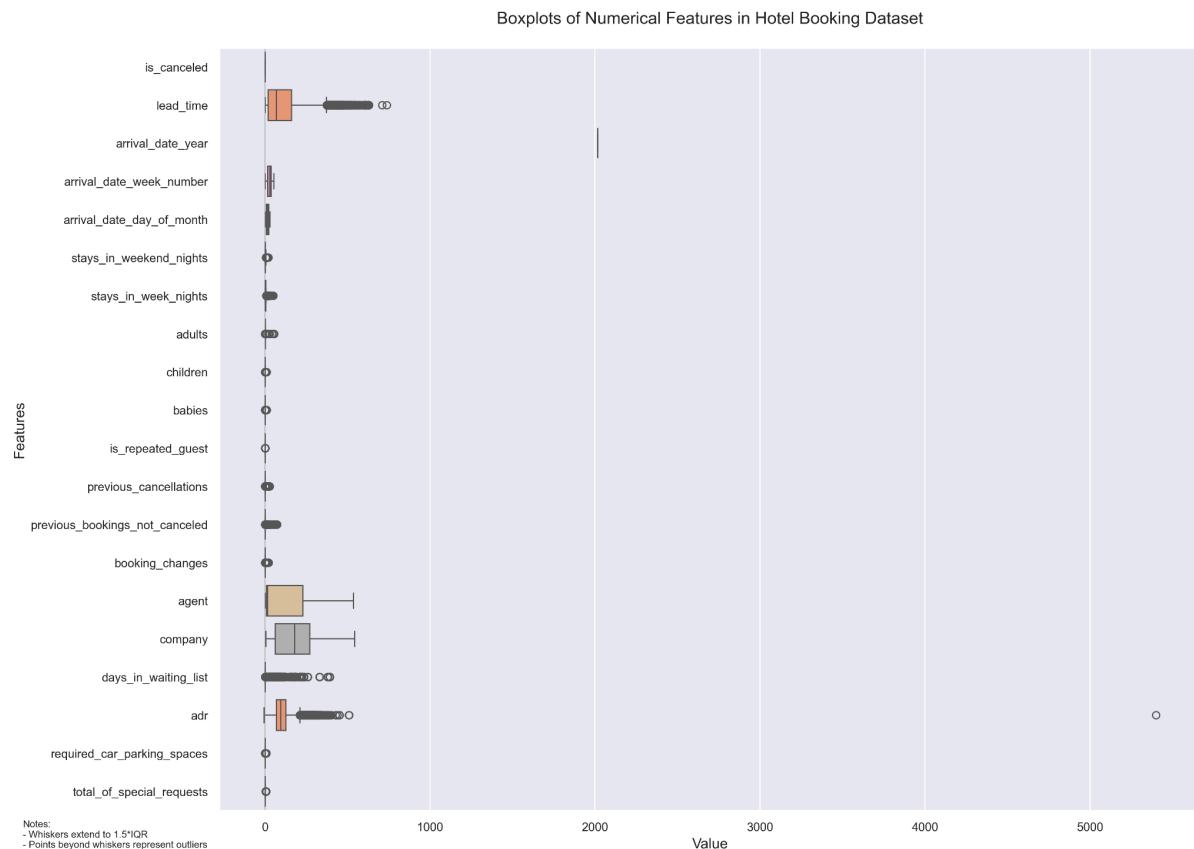
Hubungan yang signifikan juga terlihat antara "arrival\_date\_year" dan "arrival\_date\_week\_number" dengan korelasi 0,54, menunjukkan bahwa minggu kedatangan berkorelasi dengan tahun kedatangan, yang kemungkinan mencerminkan pola musiman dalam dataset multi-tahun. Korelasi positif moderat (0,42) juga ditemukan antara "is\_repeated\_guest" dan "previous\_bookings\_not\_canceled", mengkonfirmasi bahwa tamu yang pernah menginap sebelumnya dan tidak membatalkan pemesanan cenderung kembali sebagai tamu berulang.

Terkait dengan fitur target "is\_canceled", korelasi tertinggi adalah dengan "lead\_time" (0,29), mengindikasikan bahwa pemesanan dengan jarak waktu lebih panjang antara tanggal pemesanan dan kedatangan memiliki kecenderungan pembatalan yang lebih tinggi. Korelasi negatif antara "is\_canceled" dan "required\_car\_parking\_spaces" (-0,20) serta "total\_of\_special\_requests" (-0,23) menunjukkan bahwa tamu yang membuat permintaan khusus atau memerlukan tempat parkir cenderung tidak membatalkan pemesanan mereka, kemungkinan karena mereka lebih berkomitmen terhadap kunjungan yang direncanakan.

Hubungan positif moderat (0,35) antara "agent" dan "company" mengindikasikan bahwa pemesanan melalui agen tertentu sering dikaitkan dengan perusahaan tertentu, menunjukkan kemungkinan adanya kerjasama bisnis. Korelasi positif antara "adr" (Average Daily Rate) dan "children" (0,32) mengisyaratkan bahwa keluarga dengan anak-anak cenderung memesan kamar dengan tarif lebih tinggi, mungkin karena membutuhkan akomodasi yang lebih besar atau fasilitas tambahan.

Secara keseluruhan, korelasi dalam dataset ini cenderung lemah hingga moderat, dengan mayoritas nilai di bawah 0,3, mengindikasikan bahwa sebagian besar fitur memiliki hubungan independen dan memberikan informasi unik yang dapat bermanfaat untuk pemodelan prediktif. Fitur-fitur seperti "arrival\_date\_day\_of\_month" dan "babies" menunjukkan korelasi sangat rendah dengan fitur lainnya, menandakan bahwa mereka mungkin kurang berpengaruh dalam pola pemesanan hotel atau pembatalan.

## 2.2.6. Analisis Outlier



Berdasarkan visualisasi boxplot dari fitur numerik dalam dataset Hotel Booking Demand, terlihat beberapa pola penting terkait distribusi data dan keberadaan outlier yang perlu diperhatikan dalam proses preprocessing data.

Fitur "lead\_time" menunjukkan distribusi yang miring ke kanan (right-skewed) dengan banyak outlier di atas kuartil ketiga, mengindikasikan bahwa sebagian besar pemesanan dilakukan dalam jangka waktu pendek hingga menengah sebelum tanggal kedatangan, namun terdapat sejumlah kasus dengan waktu pemesanan yang sangat jauh dari hari check-in (hingga lebih dari 1000 hari). Fitur "adr" (Average Daily Rate) juga menampilkan distribusi yang sangat miring dengan outlier ekstrem mencapai nilai sekitar 5000, yang jauh melampaui nilai median. Outlier pada tarif kamar ini kemungkinan merepresentasikan pemesanan untuk suite premium atau kamar khusus pada masa peak season.

Pola serupa juga terlihat pada "days\_in\_waiting\_list" dengan mayoritas nilai mendekati nol tapi dengan beberapa outlier yang signifikan, menunjukkan bahwa sebagian besar pemesanan langsung dikonfirmasi, sementara beberapa kasus tertentu mengalami waktu tunggu yang panjang. Kolom "agent" dan "company" juga menunjukkan distribusi yang tidak seragam dengan banyak outlier, yang kemungkinan mencerminkan pola bisnis di mana beberapa agen atau perusahaan memiliki kode ID lebih tinggi dalam sistem pemesanan.

Fitur kategorikal biner seperti "is\_canceled", "is\_repeated\_guest", dan "required\_car\_parking\_spaces" menampilkan distribusi yang sangat terkonsentrasi pada nilai 0 dan 1, sesuai dengan karakteristik data biner. Sementara itu, variabel yang berkaitan dengan jumlah tamu ("adults", "children", "babies") menunjukkan distribusi yang relatif terpusat dengan outlier minimal, menandakan bahwa kebanyakan pemesanan memiliki jumlah tamu yang konsisten dalam kisaran normal.

Fitur waktu seperti "arrival\_date\_year", "arrival\_date\_week\_number", dan "arrival\_date\_day\_of\_month" memiliki distribusi yang relatif seragam dalam range yang diharapkan, dengan outlier minimal, yang mencerminkan distribusi alami dari data temporal. Variabel seperti "stays\_in\_weekend\_nights" dan "stays\_in\_week\_nights" menunjukkan pola distribusi yang miring ke kanan dengan sebagian besar nilai terkonsentrasi pada angka kecil (0-2 malam), namun terdapat outlier yang mengindikasikan pemesanan jangka panjang.

Dalam konteks preprocessing, outlier pada "lead\_time", "days\_in\_waiting\_list", dan "adr" memerlukan perhatian khusus. Meskipun outlier ini merepresentasikan kasus bisnis yang valid (bukan kesalahan data), transformasi logaritmik atau winsorization dapat dipertimbangkan untuk mengurangi dampaknya pada pemodelan prediktif, terutama untuk algoritma yang sensitif terhadap distribusi yang sangat miring. Di sisi lain, outlier pada fitur-fitur lain umumnya lebih moderat dan dapat dipertahankan karena mencerminkan variabilitas natural dalam perilaku pemesanan hotel.

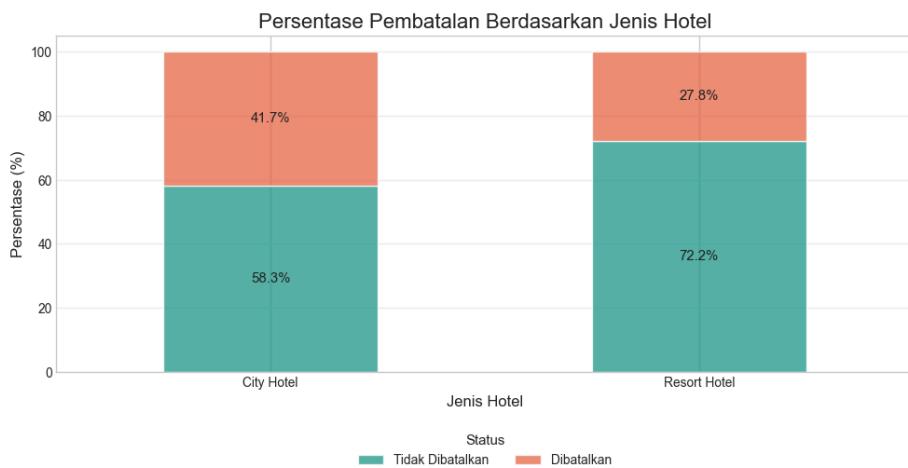
## 2.2.7. Distribusi Pembatalan Reservasi Hotel



Visualisasi pada gambar diatas menunjukkan distribusi pembatalan reservasi pada dataset hotel booking demand dengan pola yang signifikan. Dari total 119.390 reservasi, sebanyak 75.166 reservasi (63,0%) berhasil diselesaikan tanpa pembatalan, sementara 44.224 reservasi (37,0%) mengalami pembatalan. Tingkat pembatalan yang mencapai lebih dari sepertiga dari total reservasi ini mengindikasikan adanya potensi perbaikan operasional yang substansial bagi manajemen hotel, karena setiap pembatalan berpotensi mengurangi pendapatan dan efisiensi penggunaan kapasitas. Pola distribusi ini menekankan pentingnya memahami faktor-faktor yang menyebabkan pembatalan dan mengembangkan strategi untuk menurunkan angka tersebut, seperti kebijakan deposit yang lebih ketat, insentif untuk

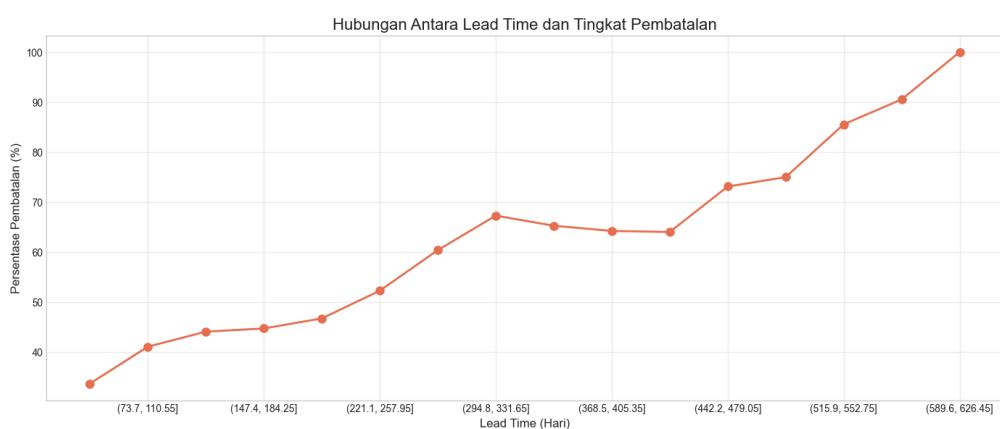
reservasi yang tidak dibatalkan, atau analisis prediktif untuk mengidentifikasi reservasi dengan risiko pembatalan tinggi sejak awal proses pemesanan.

### 2.2.8. Tingkat Pembatalan Berdasarkan Jenis Hotel



Visualisasi pada gambar diatas mengungkapkan perbedaan pola pembatalan signifikan antara kedua jenis hotel, dengan city hotel mengalami tingkat pembatalan 41,7% yang jauh lebih tinggi dibandingkan resort hotel yang hanya 27,8%. Kesenjangan sebesar 13,9% ini menunjukkan bahwa tamu city hotel memiliki kecenderungan lebih besar untuk mengubah rencana perjalanan mereka, kemungkinan karena karakteristik kunjungan bisnis atau kota yang lebih fleksibel, dibandingkan dengan tamu resort yang biasanya merencanakan liburan dengan lebih matang dan jangka waktu lebih panjang. Temuan ini memberikan masukan penting bagi manajemen city hotel untuk mempertimbangkan kebijakan pembatalan yang lebih ketat dibandingkan resort hotel.

### 2.2.9. Hubungan Lead Time dan Tingkat Pembatalan Hotel



Grafik ini memperlihatkan korelasi positif yang sangat kuat antara lead time dan persentase pembatalan, dengan pola yang konsisten menunjukkan bahwa semakin panjang lead time, semakin tinggi tingkat pembatalan. Tingkat pembatalan meningkat dari sekitar 34% pada lead time terpendek (73,7-110,55 hari) hingga mencapai 100% pada lead time terpanjang (589,6-626,45 hari), dengan kenaikan yang relatif stabil pada rentang lead time

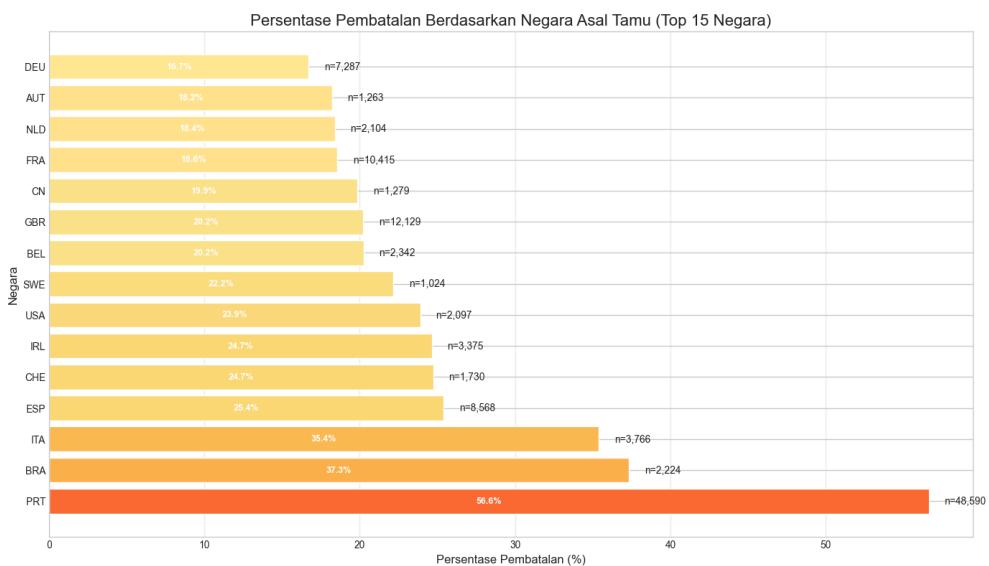
menengah dan peningkatan drastis setelah 400 hari. Pola ini mempertegas bahwa pemesanan yang dilakukan jauh-jauh hari sebelum check-in memiliki risiko pembatalan yang jauh lebih tinggi, kemungkinan karena lebih banyak faktor yang dapat berubah dalam rencana perjalanan pelanggan selama periode tunggu yang panjang. Temuan ini menyarankan kepada manajemen hotel untuk menerapkan sistem deposit progresif atau konfirmasi berkala pada pemesanan dengan lead time lebih dari 400 hari, serta memberikan insentif khusus bagi tamu yang melakukan pemesanan dengan lead time lebih pendek.

### **2.2.10. Pola Musiman Pembatalan Berdasarkan Bulan Kedatangan**



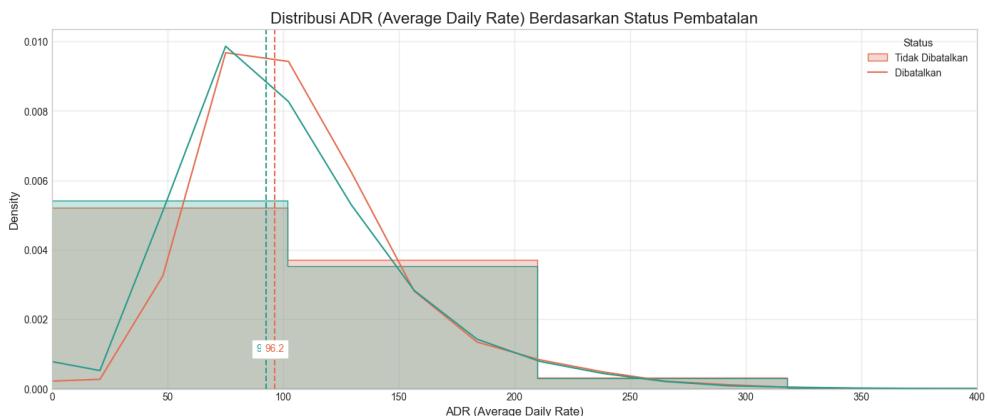
Visualisasi ini menunjukkan variasi yang signifikan dalam tingkat pembatalan hotel berdasarkan bulan kedatangan tamu, dengan pola musiman yang jelas teridentifikasi. Periode April hingga Oktober konsisten menunjukkan tingkat pembatalan yang tinggi (berkisar 37,5%-41,5%), dengan puncak pembatalan tertinggi terjadi pada bulan Juni (41,5%) dan April (40,8%), yang bertepatan dengan musim liburan musim panas di belahan bumi utara. Sebaliknya, bulan-bulan musim dingin seperti November (31,2%), Januari (30,5%), dan Maret (32,2%) mencatat tingkat pembatalan yang relatif lebih rendah, dengan selisih hingga 11% dibandingkan puncak musim. Pola ini mengindikasikan bahwa reservasi untuk periode high season pariwisata lebih rentan mengalami pembatalan, kemungkinan karena faktor kompetisi harga antar properti dan perubahan rencana liburan yang lebih sering terjadi pada masa-masa tersebut, sehingga memerlukan strategi pengelolaan kapasitas dan kebijakan pembatalan yang lebih adaptif pada bulan-bulan dengan tingkat pembatalan tinggi.

### 2.2.11. Pembatalan Berdasarkan Negara Asal Tamu



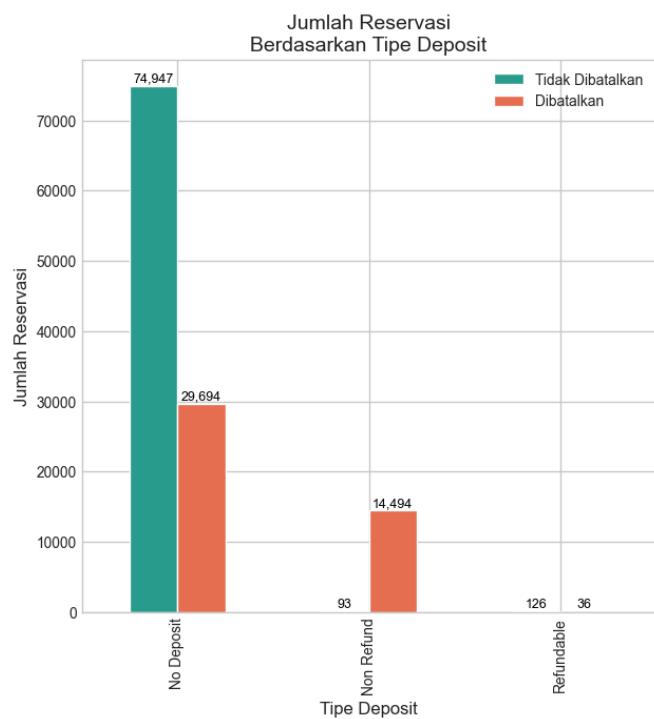
Visualisasi ini mengungkapkan disparitas yang signifikan dalam pola pembatalan hotel berdasarkan negara asal tamu. Portugal (PRT) sebagai negara tuan rumah mencatat tingkat pembatalan tertinggi (56,6%) dengan jumlah reservasi yang juga paling dominan (n=48.590), jauh melebihi negara-negara lainnya. Brasil (37,3%) dan Italia (35,4%) menempati posisi kedua dan ketiga dengan tingkat pembatalan yang moderat. Sementara itu, negara-negara Eropa Barat dan Utara seperti Jerman (16,7%), Austria (18,2%), dan Belanda (18,4%) menunjukkan tingkat pembatalan yang relatif rendah. Perbedaan signifikan antara tamu domestik dan internasional ini mungkin dipengaruhi oleh perbedaan perilaku pemesanan, di mana tamu lokal cenderung lebih fleksibel dalam membuat dan membatalkan reservasi karena jarak yang lebih dekat dan biaya perjalanan yang lebih rendah. Temuan ini mengisyaratkan pentingnya strategi pemasaran dan kebijakan pembatalan yang disesuaikan berdasarkan geografis, dengan penekanan khusus pada upaya retensi tamu domestik yang memiliki tingkat pembatalan tertinggi.

### 2.2.12. Distribusi ADR Berdasarkan Status Pembatalan



Visualisasi ini menunjukkan distribusi Average Daily Rate (ADR) berdasarkan status pembatalan dengan pola yang cukup informatif. Kedua kurva distribusi memiliki bentuk yang serupa dengan puncak di kisaran €75-€100, namun terlihat perbedaan subtil di mana reservasi yang dibatalkan (garis merah) memiliki rata-rata ADR yang sedikit lebih tinggi (€96,2) dibandingkan dengan yang tidak dibatalkan (garis hijau, €94,5). Distribusi untuk reservasi yang dibatalkan juga menunjukkan densitas yang lebih tinggi di rentang harga €100-€150, mengindikasikan bahwa reservasi dengan tarif lebih tinggi memiliki kecenderungan pembatalan yang lebih besar. Fenomena ini mungkin mencerminkan perilaku konsumen di mana pemesanan dengan harga lebih mahal lebih sering dievaluasi ulang oleh tamu, kemungkinan karena pertimbangan nilai ekonomis atau pencarian alternatif yang lebih murah seiring mendekati tanggal kedatangan. Temuan ini menyarankan bahwa hotel perlu menerapkan strategi retensi yang lebih kuat untuk pemesanan dalam segmen harga menengah ke tinggi, seperti penawaran nilai tambah atau insentif loyalitas untuk mengurangi risiko pembatalan.

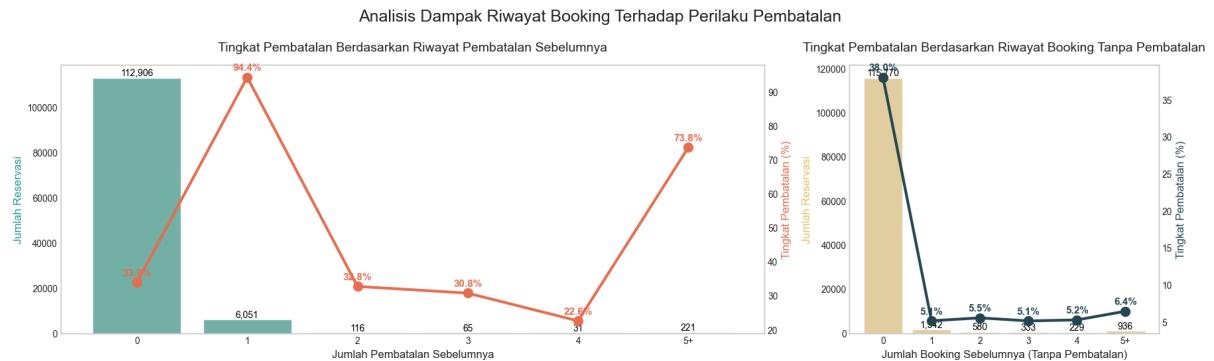
### 2.2.13. Hubungan Tipe Deposit dengan Tingkat Pembatalan



Visualisasi ini mengungkapkan pengaruh signifikan kebijakan deposit terhadap perilaku pembatalan reservasi hotel. Kategori "No Deposit" mendominasi jumlah reservasi dengan total 104.641 pemesanan, namun memiliki tingkat pembatalan moderat (28,4%). Yang sangat kontras adalah kategori "Non Refund" dengan 14.587 reservasi, di mana hampir seluruhnya (99,4%) berakhir dengan pembatalan, menunjukkan bahwa kebijakan deposit yang tidak dapat dikembalikan justru menjadi indikator kuat untuk pembatalan. Sebaliknya, opsi "Refundable" meskipun dengan jumlah sangat kecil (162 reservasi) memiliki tingkat pembatalan terendah (22,2%), mengindikasikan bahwa fleksibilitas pengembalian dana cenderung menurunkan risiko pembatalan. Pola ini menunjukkan paradoks menarik dalam

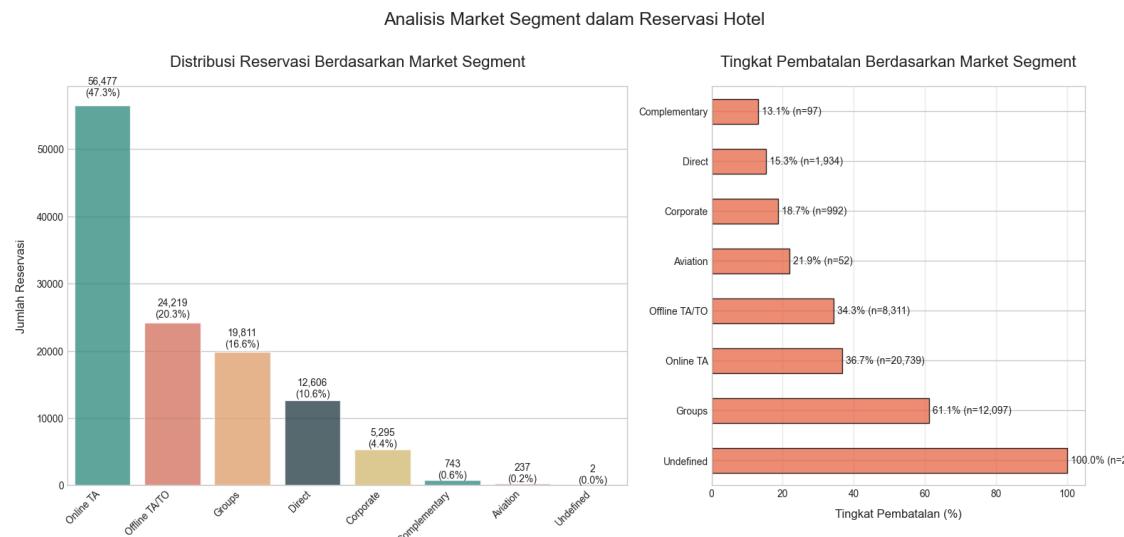
preferensi konsumen hotel, di mana kebijakan deposit yang dirancang untuk mencegah pembatalan (non-refundable) justru berkorelasi dengan tingkat pembatalan tertinggi, kemungkinan karena tamu merasa terjebak dengan pilihan yang tidak fleksibel atau mengalami perubahan rencana yang tidak dapat diakomodasi.

#### 2.2.14. Dampak Riwayat Booking Terhadap Perilaku Pembatalan



Visualisasi ini mengungkapkan hubungan yang kontras antara riwayat pembatalan dan riwayat booking sukses terhadap perilaku pembatalan. Pada grafik kiri, terdapat korelasi positif yang kuat antara jumlah pembatalan sebelumnya dengan tingkat pembatalan saat ini; tamu dengan satu pembatalan sebelumnya menunjukkan tingkat pembatalan yang sangat tinggi (94,4%), jauh melebihi tamu tanpa riwayat pembatalan (36,7%), dan pola ini berlanjut dengan tingkat pembatalan 73,8% untuk tamu dengan 5+ pembatalan sebelumnya. Sebaliknya, grafik kanan memperlihatkan hubungan negatif yang jelas antara jumlah booking sukses dengan tingkat pembatalan; tamu tanpa riwayat booking sukses memiliki tingkat pembatalan tertinggi (37,5%), sementara tamu dengan satu atau lebih booking sukses sebelumnya menunjukkan tingkat pembatalan yang signifikan lebih rendah (di bawah 6,4%). Temuan ini menggarisbawahi pentingnya segmentasi pelanggan berdasarkan histori pemesanan, dengan implikasi bahwa tamu dengan riwayat pembatalan merupakan segmen berisiko tinggi yang memerlukan strategi retensi khusus, sementara tamu yang telah menyelesaikan kunjungan sebelumnya cenderung menjadi pelanggan yang lebih loyal dan dapat diandalkan.

## 2.2.15. Market Segment dalam Reservasi Hotel



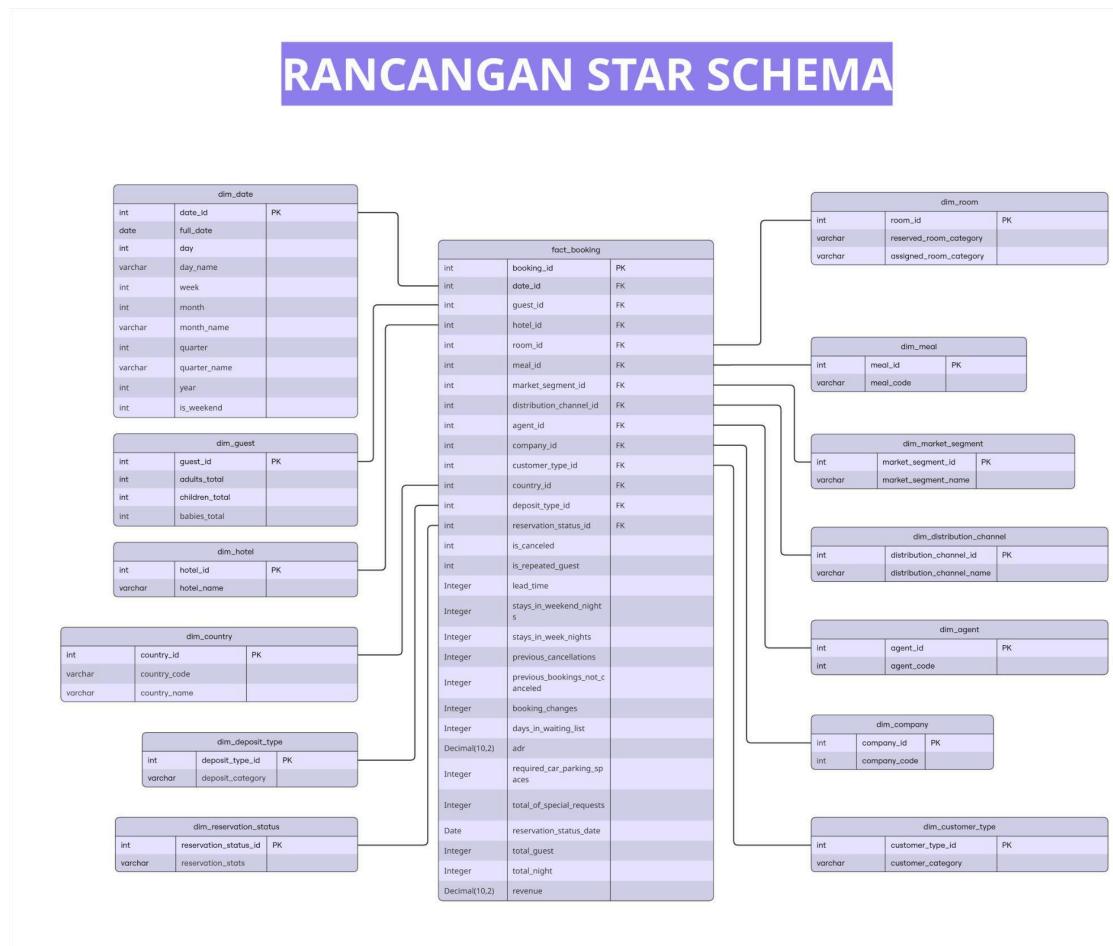
Visualisasi ini mengungkapkan pola distribusi dan pembatalan yang kontras di antara berbagai segmen pasar hotel. Segmen Online TA (Travel Agency) mendominasi volume reservasi dengan 56.477 pemesanan (47,3% dari total), diikuti oleh Offline TA/TO (20,3%) dan Groups (16,6%), menunjukkan ketergantungan signifikan industri pada saluran distribusi online. Namun, yang menarik adalah hubungan terbalik antara volume pemesanan dan tingkat pembatalan—segmen Groups memiliki tingkat pembatalan tertinggi di antara segmen utama (61,1%), jauh di atas Online TA (36,7%) dan Offline TA/TO (34,3%). Sementara itu, segmen korporat dan pemesanan langsung (Direct) menunjukkan loyalitas lebih tinggi dengan tingkat pembatalan masing-masing hanya 18,7% dan 15,3%, meskipun volumenya lebih rendah. Pola ini menunjukkan bahwa segmen pasar dengan volume tinggi dan harga terendah (seperti Groups dan Online TA) menghasilkan lebih banyak pemesanan tetapi juga risiko pembatalan yang lebih tinggi, sementara segmen seperti Corporate and Direct mungkin menghasilkan pendapatan lebih stabil meski dengan volume lebih rendah, sehingga hotel perlu menyeimbangkan strategi akuisisi tamu mereka dengan mempertimbangkan tidak hanya volume tetapi juga tingkat retensi di setiap segmen.

### BAB III PERANCANGAN STAR SCHEMA

Star Schema merupakan metodologi pengorganisasian data yang efisien dalam Data Warehouse, ditandai dengan satu tabel fakta sebagai pusat yang terhubung dengan beberapa tabel dimensi di sekelilingnya. Struktur ini menyerupai bentuk bintang (star), sehingga memberikan keunggulan berupa desain yang intuitif, performa query yang optimal, dan kemudahan dalam analisis multidimensional. Dalam laporan ini, akan dibahas secara mendalam mengenai komponen-komponen utama:

- Tabel Fakta (Fact Table): Merupakan pusat dari star schema yang berisi data transaksional utama beserta foreign key yang merujuk ke tabel dimensi. Tabel ini menyimpan metrics dan measures yang menjadi fokus analisis bisnis.
- Tabel Dimensi (Dimension Table): Mengandung atribut deskriptif yang memperkaya konteks data pada tabel fakta. Tabel ini berfungsi sebagai referensi untuk mengategorikan dan memfilter data dalam proses analisis.
- Measure: Representasi kuantitatif dari suatu proses bisnis yang digunakan untuk mengukur performa. Measures dapat berupa nilai yang tersimpan langsung dalam tabel fakta atau nilai turunan yang dihitung dari beberapa field.

Berikut merupakan rancangan Star Schema pada tugas besar ini untuk sistem hotel booking, yang terdiri dari tabel fakta (fact\_booking) dan beberapa tabel dimensi.



Link Star Schema

### 3.1. Fact Table (fact\_booking)

Tabel fakta adalah pusat dari Star Schema, yang menyimpan data transaksi pemesanan hotel utama. Tabel ini berisi nilai-nilai numerik yang dapat diukur (measure) dan hubungan dengan tabel dimensi melalui foreign key (FK).

Kolom	Tipe Data	Deskripsi
booking_id (PK)	INT	ID unik untuk setiap pemesanan hotel
date_id (FK)	INT	Referensi ke tabel dim_date
guest_id (FK)	INT	Referensi ke tabel dim_guest
hotel_id (FK)	INT	Referensi ke tabel dim_hotel
room_id (FK)	INT	Referensi ke tabel dim_room
meal_id (FK)	INT	Referensi ke tabel dim_meal
market_segment_id (FK)	INT	Referensi ke tabel dim_market_segment
distribution_channel_id (FK)	INT	Referensi ke tabel dim_distribution_channel
agent_id (FK)	INT	Referensi ke tabel dim_agent, dapat NULL jika tidak melalui agen
company_id (FK)	INT	Referensi ke tabel dim_company, dapat NULL jika bukan pemesanan perusahaan
customer_type_id (FK)	INT	Referensi ke tabel dim_customer_type
country_id (FK)	INT	Referensi ke tabel dim_country
deposit_type_id (FK)	INT	Referensi ke tabel dim_deposit_type
reservation_status_id (FK)	INT	Referensi ke tabel dim_reservation_status
is_canceled	BOOLEAN	Status apakah pemesanan dibatalkan (1) atau tidak (0)
is_repeated_guest	BOOLEAN	Status apakah tamu merupakan tamu berulang (1) atau baru (0)

lead_time	INT	Jumlah hari antara tanggal pemesanan dan tanggal kedatangan
stays_in_weekend_nights	INT	Jumlah menginap di malam akhir pekan (Sabtu/Minggu)
stays_in_week_nights	INT	Jumlah menginap di malam hari kerja (Senin-Jumat)
previous_cancellations	INT	Jumlah pembatalan sebelumnya yang dilakukan tamu
previous_bookings_not_canceled	INT	Jumlah pemesanan sebelumnya yang tidak dibatalkan oleh tamu
booking_changes	INT	Jumlah perubahan/amandemen yang dilakukan pada pemesanan
days_in_waiting_list	INT	Jumlah hari pemesanan berada dalam daftar tunggu
adr	DECIMAL(10,2)	Average Daily Rate - Rata-rata harga per malam selama masa tinggal
required_car_parking_spaces	INTEGER	Jumlah tempat parkir mobil yang diminta oleh tamu
total_of_special_requests	INTEGER	Jumlah permintaan khusus dari tamu
reservation_status_date	DATE	Tanggal status reservasi terakhir diperbarui
total_guest	INTEGER	(Turunan) Jumlah total tamu (adults_total + children_total + babies_total)
total_night	INTEGER	(Turunan) Jumlah total malam menginap (stays_in_weekend_nights + stays_in_week_nights)
revenue	DECIMAL(10,2)	(Turunan) Total pendapatan (adr * total_night)

Fungsi Tabel Fakta:

1. Analisis Pemesanan Hotel:

- a. Memungkinkan analisis jumlah pemesanan berdasarkan berbagai dimensi seperti waktu, lokasi, jenis pelanggan, dsb.
  - b. Mendukung perhitungan tingkat pembatalan dan faktor-faktor yang mempengaruhinya.
2. Revenue Analysis:
- a. Memungkinkan perhitungan pendapatan berdasarkan berbagai segmen pasar, saluran distribusi, dan jenis kamar.
  - b. Mendukung analisis ADR (Average Daily Rate) dan revenue per available room (RevPAR).
3. Customer Behavior Analysis:
- a. Memungkinkan analisis perilaku pelanggan seperti pola menginap, kebutuhan khusus, dan loyalitas.
  - b. Mendukung segmentasi pelanggan berdasarkan karakteristik pemesanan mereka.
4. Operational Efficiency Metrics:
- a. Menyediakan data untuk menganalisis efisiensi operasional hotel seperti waktu tunggu pemesanan dan perubahan pemesanan.

### 3.2. Dimension Table

Tabel dimensi digunakan untuk menyimpan informasi deskriptif terkait data transaksi yang ada di tabel fakta.

a. dim\_date

Tabel dimensi waktu menyimpan informasi terkait tanggal untuk mendukung analisis temporal. Tabel ini memungkinkan agregasi data pada berbagai tingkat granularitas waktu (hari, minggu, bulan, kuartal, tahun).

Kolom	Tipe Data	Deskripsi
date_id (PK)	INT	ID unik yang mengidentifikasi tanggal
full_date	DATE	Tanggal lengkap dalam format date
day	INT	Hari dalam bulan (1-31)
day_name	VARCHAR(10)	Nama hari (Senin, Selasa, dst)
week	INT	Nomor minggu dalam tahun (1-53)
month	INT	Nomor bulan (1-12)
month_name	VARCHAR(10)	Nama bulan (Januari, Februari, dst)
quarter	INT	Kuartal dalam tahun (1-4)

quarter_name	VARCHAR(10)	Nama kuartal (Q1, Q2, Q3, Q4)
year	INT	Tahun
is_weekend	BOOLEAN	Penanda apakah tanggal tersebut adalah akhir pekan (TRUE/FALSE)

b. dim\_guest

Tabel dimensi guest menyimpan informasi terkait tamu hotel, termasuk komposisi tamu dalam reservasi.

Kolom	Tipe Data	Deskripsi
guest_id (PK)	INT	ID unik yang mengidentifikasi tamu
adults_total	INT	Jumlah tamu dewasa
children_total	INT	Jumlah anak-anak
babies_total	INT	Jumlah bayi dalam reservasi

c. dim\_hotel

Tabel dimensi hotel menyimpan informasi tentang properti hotel yang tersedia dalam sistem.

Kolom	Tipe Data	Deskripsi
hotel_id (PK)	INT	ID unik yang mengidentifikasi hotel
hotel_name	VARCHAR(100)	Nama hotel (Resort Hotel, City Hotel, dll)

d. dim\_room

Tabel dimensi room menyimpan informasi tentang tipe kamar hotel dan kategorinya.

Kolom	Tipe Data	Deskripsi
room_id (PK)	INT	ID unik yang mengidentifikasi kamar
reserved_room_category	VARCHAR(50)	Kategori kamar yang direservasi (A, B, C, D, E, F, G, H, dll)

assigned_room_category	VARCHAR(50)	Kategori kamar yang diberikan (dapat berbeda dari yang direservasi)
------------------------	-------------	---

e. dim\_meal

Tabel dimensi meal menyimpan informasi tentang paket makanan yang ditawarkan hotel.

Kolom	Tipe Data	Deskripsi
meal_id (PK)	INT	ID unik yang mengidentifikasi paket makanan
meal_code	VARCHAR(50)	Kode paket makanan (BB, HB, FB, dll) - Breakfast, Half Board, Full Board

f. dim\_market\_segment

Tabel dimensi market\_segment menyimpan informasi tentang segmen pasar yang menjadi sumber pemesanan.

Kolom	Tipe Data	Deskripsi
market_segment_id (PK)	INT	ID unik yang mengidentifikasi segmen pasar
market_segment_name	VARCHAR(50)	Nama segmen pasar (Direct, Corporate, Online TA, Offline TA/TO, Groups, dll)

g. dim\_distribution\_channel

Tabel dimensi distribution\_channel menyimpan informasi tentang saluran distribusi yang digunakan untuk pemesanan.

Kolom	Tipe Data	Deskripsi
distribution_channel_id (PK)	INT	ID unik yang mengidentifikasi saluran distribusi
distribution_channel_name	VARCHAR(50)	Nama saluran distribusi (Direct, Corporate, TA/TO, dll)

h. dim\_agent

Tabel dimensi agent menyimpan informasi tentang agen travel yang melakukan pemesanan.

Kolom	Tipe Data	Deskripsi
agent_id (PK)	INT	ID unik yang mengidentifikasi agen travel
agent_code	INT	Kode internal agen travel

i. dim\_company

Tabel dimensi company menyimpan informasi tentang perusahaan yang melakukan pemesanan korporat.

Kolom	Tipe Data	Deskripsi
company_id (PK)	INT	ID unik yang mengidentifikasi perusahaan
company_code	INT	Kode internal perusahaan

j. dim\_customer\_type

Tabel dimensi customer\_type menyimpan informasi tentang tipe pelanggan yang melakukan pemesanan.

Kolom	Tipe Data	Deskripsi
customer_type_id (PK)	INT	ID unik yang mengidentifikasi tipe pelanggan
customer_category	VARCHAR(50)	Kategori pelanggan (Transient, Contract, Transient-Party, Group, dll)

k. dim\_country

Tabel dimensi country menyimpan informasi tentang negara asal tamu hotel.

Kolom	Tipe Data	Deskripsi
country_id (PK)	INT	ID unik yang mengidentifikasi negara

country_code	VARCHAR(100)	Kode negara berdasarkan standar ISO (PRT, ESP, GBR, dll)
country_name	VARCHAR(100)	Nama lengkap negara (Portugal, Spain, United Kingdom, dll)

l. dim\_deposit\_type

Tabel dimensi deposit\_type menyimpan informasi tentang jenis deposit yang digunakan dalam pemesanan.

Kolom	Tipe Data	Deskripsi
deposit_type_id (PK)	INT	ID unik yang mengidentifikasi jenis deposit
deposit_category	VARCHAR(50)	Kategori deposit (No Deposit, Non-Refundable, Refundable, dll)

m. dim\_reservation\_status

Tabel dimensi reservation\_status menyimpan informasi tentang status akhir dari reservasi.

Kolom	Tipe Data	Deskripsi
reservation_statuses_id (PK)	INT	ID unik yang mengidentifikasi status reservasi
reservation_statuses	VARCHAR(50)	Status reservasi (Confirmed, Checked-Out, Canceled, No-Show, dll)

### 3.3. Measure dalam Analisis Data Warehouse Hotel Booking

Measure adalah nilai kuantitatif yang dapat dihitung dan dianalisis dalam bisnis perhotelan. Measure ini digunakan untuk mengukur performa operasional hotel, memahami pola pemesanan, dan membantu dalam pengambilan keputusan strategis untuk meningkatkan pendapatan dan kepuasan pelanggan. Berikut adalah daftar Measure dalam Analisis Hotel Booking:

Measure	Formula	Deskripsi

total_guest	adults_total + children_total + babies_total	Total jumlah tamu dalam setiap pemesanan. Mengukur volume okupansi dan membantu dalam perencanaan kapasitas dan layanan hotel.
total_nights	stays_in_weekend_nights + stays_in_week_nights	Jumlah total malam menginap yang dipesan. Metrik ini penting untuk perhitungan pendapatan dan analisis durasi kunjungan.
revenue	adr * total_nights	Total pendapatan yang dihasilkan dari pemesanan. Merupakan indikator kunci performa finansial dari setiap transaksi hotel.

## BAB IV IMPLEMENTASI STAR SCHEMA DALAM RDBMS

Dalam implementasi ini, kelompok kami mendefinisikan struktur tabel dengan menentukan nama database, tabel (fact dan dimension table), kolom pada masing-masing tabel, tipe data, serta menetapkan primary key dan foreign key antar tabel sesuai dengan struktur rancangan star schema.

### 4.1. Pembuatan Database

#### Syntax DDL MySQL

```
CREATE DATABASE IF NOT EXISTS `hotel_dwbi` DEFAULT CHARACTER SET  
utf8mb4 COLLATE utf8mb4_0900_ai_ci;  
USE `hotel_dwbi`;
```

### 4.2. Pembuatan Dimension Table

Syntax DDL MySQL	Deskripsi
-- Membuat tabel fakta booking (fact_booking) CREATE TABLE fact_booking ( booking_id INT PRIMARY KEY, date_id INT NOT NULL, guest_id INT NOT NULL, hotel_id INT NOT NULL, room_id INT NOT NULL, meal_id INT NOT NULL, market_segment_id INT NOT NULL, distribution_channel_id INT NOT NULL, agent_id INT NULL, company_id INT NULL, customer_type_id INT NOT NULL, country_id INT NOT NULL, deposit_type_id INT NOT NULL, reservation_status_id INT NOT NULL, is_canceled INT NOT NULL, is_repeated_guest INT NOT NULL, lead_time INT NOT NULL, stays_in_weekend_nights INT NOT NULL, stays_in_week_nights INT NOT NULL, previous_cancellations INT NOT NULL, previous_bookings_not_canceled INT NOT NULL,	Membuat tabel pusat (fact table) yang menyimpan informasi transaksi pemesanan hotel dengan berbagai atribut penting seperti status pembatalan, lama menginap, dan metrik finansial serta menghubungkan ke seluruh tabel dimensi melalui foreign key untuk memungkinkan analisis multi-dimensi.

```

booking_changes INT NOT NULL,
days_in_waiting_list INT NOT NULL,
adr DECIMAL(10,2) NOT NULL,
required_car_parking_spaces INT NOT NULL,
total_of_special_requests INT NOT NULL,
reservation_status_date DATE NOT NULL,
total_guest INT NOT NULL,
total_night INT NOT NULL,
revenue DECIMAL(10,2) NOT NULL,
FOREIGN KEY (date_id) REFERENCES
dim_date(date_id),
FOREIGN KEY (guest_id) REFERENCES
dim_guest(guest_id),
FOREIGN KEY (hotel_id) REFERENCES
dim_hotel(hotel_id),
FOREIGN KEY (room_id) REFERENCES
dim_room(room_id),
FOREIGN KEY (meal_id) REFERENCES
dim_meal(meal_id),
FOREIGN KEY (market_segment_id)
REFERENCES
dim_market_segment(market_segment_id),
FOREIGN KEY (distribution_channel_id)
REFERENCES
dim_distribution_channel(distribution_channel_id),
FOREIGN KEY (agent_id) REFERENCES
dim_agent(agent_id),
FOREIGN KEY (company_id) REFERENCES
dim_company(company_id),
FOREIGN KEY (customer_type_id)
REFERENCES
dim_customer_type(customer_type_id),
FOREIGN KEY (country_id) REFERENCES
dim_country(country_id),
FOREIGN KEY (deposit_type_id)
REFERENCES
dim_deposit_type(deposit_type_id),
FOREIGN KEY (reservation_status_id)
REFERENCES
dim_reservation_status(reservation_status_id)
);

```

```

-- Indeks untuk meningkatkan performa query pada
tabel fakta
CREATE INDEX idx_fact_booking_date ON
fact_booking(date_id);
CREATE INDEX idx_fact_booking_hotel ON
fact_booking(hotel_id);
CREATE INDEX idx_fact_booking_guest ON

```

Membuat indeks pada kolom foreign key utama untuk mengoptimalkan performa query analitik yang sering digunakan, mempercepat operasi join dengan tabel dimensi, dan meningkatkan efisiensi agregasi data berdasarkan dimensi-dimensi kunci

<pre> fact_booking(guest_id); CREATE INDEX idx_fact_booking_room ON fact_booking(room_id); CREATE INDEX idx_fact_booking_meal ON fact_booking(meal_id); CREATE INDEX idx_fact_booking_market_segment ON fact_booking(market_segment_id); CREATE INDEX idx_fact_booking_distribution_channel ON fact_booking(distribution_channel_id); CREATE INDEX idx_fact_booking_country ON fact_booking(country_id); CREATE INDEX idx_fact_booking_reservation_status ON fact_booking(reservation_status_id); CREATE INDEX idx_fact_booking_deposit_type ON fact_booking(deposit_type_id); </pre>	dalam sistem data warehouse.
---	------------------------------

### 4.3. Pembuatan Dimension Table

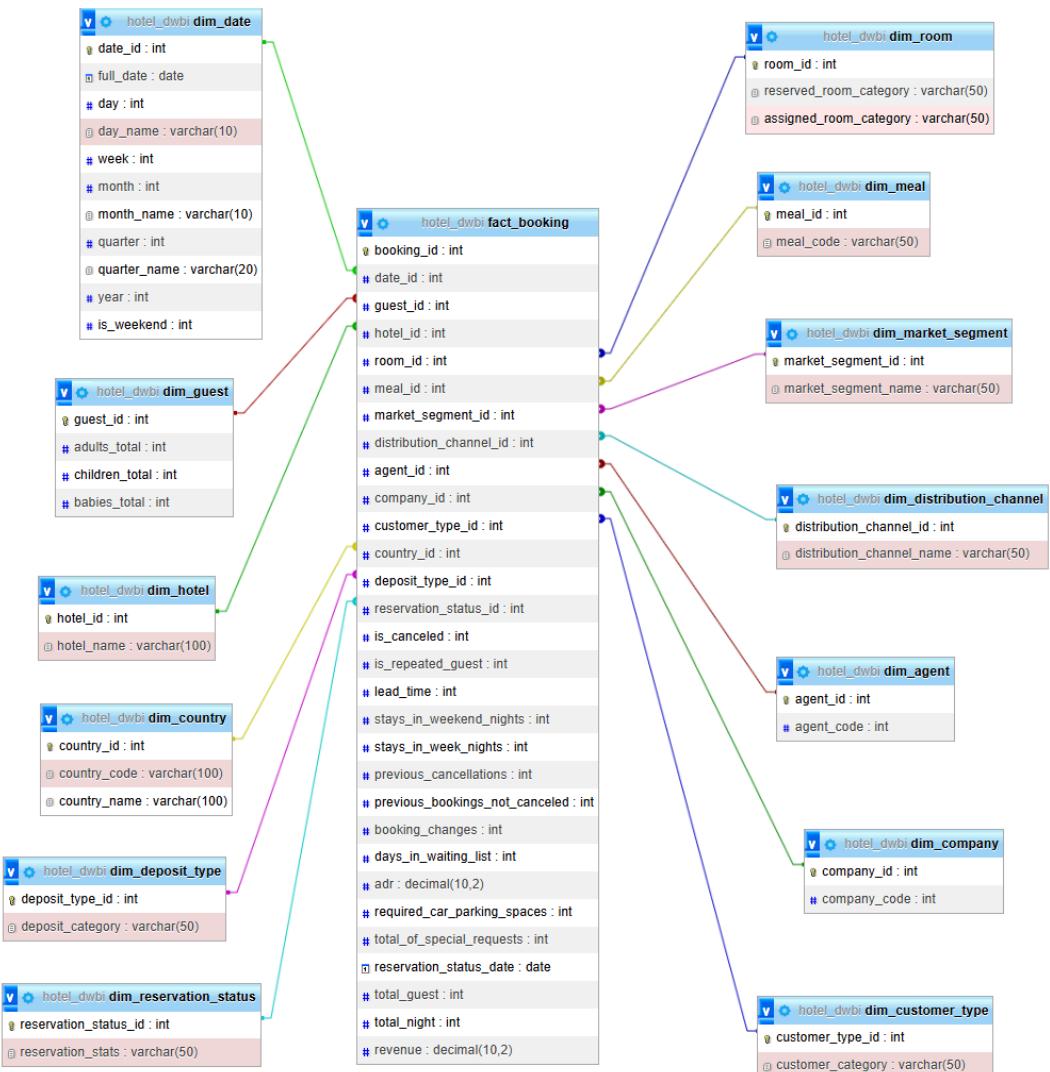
Nama Dimension Table	Syntax DDL MySQL	Deskripsi
dim_date	-- Membuat tabel dimensi tanggal (dim_date) CREATE TABLE dim_date (     date_id INT PRIMARY KEY,     full_date DATE NOT NULL,     day INT NOT NULL,     day_name VARCHAR(10) NOT NULL,     week INT NOT NULL,     month INT NOT NULL,     month_name VARCHAR(10) NOT NULL,     quarter INT NOT NULL,     quarter_name VARCHAR(20) NOT NULL,     year INT NOT NULL,     is_weekend INT NOT NULL );	Menyimpan hierarki waktu lengkap (hari, minggu, bulan, kuartal, tahun) untuk mendukung analisis temporal dan tren musiman dalam data pemesanan hotel, memungkinkan agregasi data pada berbagai level granularitas waktu.
dim_guest	-- Membuat tabel dimensi tamu (dim_guest)	Menyimpan informasi komposisi tamu dalam

	<pre>CREATE TABLE dim_guest (     guest_id INT PRIMARY KEY,     adults_total INT NOT NULL,     children_total INT NOT NULL,     babies_total INT NOT NULL );</pre>	setiap reservasi, memungkinkan analisis segmentasi berdasarkan jenis dan jumlah tamu untuk optimasi kapasitas kamar dan pengembangan penawaran paket khusus.
dim_hotel	<pre>-- Membuat tabel dimensi hotel (dim_hotel) CREATE TABLE dim_hotel (     hotel_id INT PRIMARY KEY,     hotel_name VARCHAR(100) NOT NULL );</pre>	Menyimpan data tentang properti hotel dalam jaringan, memungkinkan perbandingan performa antar hotel dan analisis strategi bisnis spesifik untuk setiap properti.
dim_country	<pre>-- Membuat tabel dimensi negara (dim_country) CREATE TABLE dim_country (     country_id INT PRIMARY KEY,     country_code VARCHAR(100) NOT NULL,     country_name VARCHAR(100) NOT NULL );</pre>	Menyimpan informasi negara asal tamu, mendukung analisis segmentasi pasar berdasarkan geografi dan membantu dalam pengembangan strategi pemasaran yang ditargetkan untuk pasar internasional yang berbeda.
dim_deposit_type	<pre>-- Membuat tabel dimensi tipe deposit (dim_deposit_type) CREATE TABLE dim_deposit_type (     deposit_type_id INT PRIMARY KEY,     deposit_category VARCHAR(50) NOT NULL );</pre>	Menyimpan jenis deposit yang digunakan dalam pemesanan (No Deposit, Non-Refundable, Refundable), memungkinkan analisis hubungan antara kebijakan deposit dengan tingkat pembatalan dan optimasi strategi pendapatan.
dim_reservation_status	<pre>-- Membuat tabel dimensi status reservasi (dim_reservation_status) CREATE TABLE dim_reservation_status (     reservation_status_id INT</pre>	Menyimpan status akhir reservasi (Checked-Out, Canceled, No-Show), mendukung analisis tren pembatalan dan pengembangan model

	<pre> PRIMARY KEY, reservation_stats VARCHAR(50) NOT NULL ); </pre>	<p>prediktif untuk mengantisipasi risiko pembatalan di masa depan.</p>
dim_room	<pre> -- Membuat tabel dimensi kamar (dim_room) CREATE TABLE dim_room (     room_id INT PRIMARY KEY,     reserved_room_category VARCHAR(50) NOT NULL,     assigned_room_category VARCHAR(50) NOT NULL ); </pre>	<p>Menyimpan informasi kategori kamar yang dipesan vs yang diberikan, memungkinkan analisis tingkat upgrade/downgrade dan optimasi strategi penetapan harga berdasarkan permintaan tipe kamar.</p>
dim_meal	<pre> -- Membuat tabel dimensi meal (dim_meal) CREATE TABLE dim_meal (     meal_id INT PRIMARY KEY,     meal_code VARCHAR(50) NOT NULL ); </pre>	<p>Menyimpan informasi paket makanan yang ditawarkan hotel (BB, HB, FB, SC), mendukung analisis preferensi paket makanan berdasarkan segmen pasar dan evaluasi dampaknya terhadap pendapatan.</p>
dim_market_segment	<pre> -- Membuat tabel dimensi segmen pasar (dim_market_segment) CREATE TABLE dim_market_segment (     market_segment_id INT PRIMARY KEY,     market_segment_name VARCHAR(50) NOT NULL ); </pre>	<p>Menyimpan segmen pasar sumber pemesanan (Direct, Corporate, Online TA, dll), memungkinkan analisis efektivitas saluran pemasaran dan pengembangan strategi penetapan harga berdasarkan segmen.</p>
dim_distribution_channel	<pre> -- Membuat tabel dimensi saluran distribusi (dim_distribution_channel) CREATE TABLE dim_distribution_channel (     distribution_channel_id INT PRIMARY KEY,     distribution_channel_name VARCHAR(50) NOT NULL ); </pre>	<p>Menyimpan saluran distribusi untuk pemesanan (Direct, Corporate, TA/TO, GDS), membantu analisis biaya akuisisi pelanggan per saluran dan identifikasi saluran yang paling menguntungkan.</p>
dim_agent	<pre> -- Membuat tabel dimensi agen (dim_agent) </pre>	<p>Menyimpan informasi agen travel yang</p>

	<pre>CREATE TABLE dim_agent (     agent_id INT PRIMARY KEY,     agent_code INT NOT NULL );</pre>	mengelola pemesanan, mendukung analisis performa penjualan per agen dan program insentif berdasarkan performa.
dim_company	<pre>-- Membuat tabel dimensi perusahaan (dim_company) CREATE TABLE dim_company (     company_id INT PRIMARY KEY,     company_code INT NOT NULL );</pre>	Menyimpan data perusahaan yang melakukan pemesanan korporat, memungkinkan analisis volume dan nilai pemesanan per perusahaan untuk negosiasi kontrak dan pengembangan program loyalitas korporat.
dim_customer_type	<pre>-- Membuat tabel dimensi tipe pelanggan (dim_customer_type) CREATE TABLE dim_customer_type (     customer_type_id INT PRIMARY KEY,     customer_category VARCHAR(50) NOT NULL );</pre>	Menyimpan tipe pelanggan (Transient, Contract, Transient-Party, Group), mendukung analisis profitabilitas berdasarkan tipe pelanggan dan optimasi strategi penetapan harga sesuai segmen.

## 4.4. Desain Struktur Database DDL (RDBMS)



[Link Star Schema RDBMS](#)

# BAB V IMPLEMENTASI PROSES ETL

## 5.1. Fact Table

Berikut merupakan implementasi berupa gambar screenshot dari proses ETL pada order\_fact yang telah dilakukan pada pentaho:



[Link HotelBooking DWBI KTR](#)

Proses Extract, Transform, Load (ETL) pada fact\_booking dalam Pentaho dilakukan pada langkah-langkah dibawah ini:

### 1. Proses Extract

Proses extract dilakukan dengan menggunakan CSV File Input baru, kemudian dilanjutkan dengan sejumlah transformasi agar data sesuai dengan struktur pada

masing-masing tabel dimensi. Dalam tahap ini, digunakan enam Value Mapper untuk memproses enam kolom, termasuk:

- Mengubah semua nilai tidak valid seperti undefined menjadi nilai modus (nilai yang paling sering muncul),
- Melakukan imputasi terhadap nilai null, misalnya pada kolom company dan agent yang diberi nilai 0 karena menunjukkan individu yang tidak diwakilkan oleh entitas tertentu.

Setiap tabel dimensi juga turut menjadi bagian dari proses extract ini, karena diperlukan dalam pembentukan fact table. Data dari masing-masing dimensi diambil dari transformasi akhir, misalnya melalui Select Values, untuk kemudian digabungkan secara bertahap menggunakan Stream Lookup. Proses ini dilakukan secara iteratif terhadap seluruh tabel dimensi untuk memastikan integrasi yang konsisten dan akurat ke dalam fact table utama.

## 2. Proses Transform

Pada tahap transformasi, seperti telah dijelaskan sebelumnya pada proses extract, dilakukan Stream Lookup untuk setiap tabel dimensi. Tujuannya adalah mencocokkan data dari sumber (CSV File Input) dengan nilai primary key dari masing-masing tabel dimensi, yang nantinya akan menjadi foreign key (FK) di dalam fact table. Meskipun istilah teknis yang digunakan bisa bervariasi, proses ini secara umum dikenal sebagai dimension key mapping atau dimensional surrogate key assignment.

Proses ini dilakukan secara iteratif untuk setiap dari 13 tabel dimensi. Artinya, data dari file CSV akan diperkaya secara bertahap — satu per satu — dengan foreign key dari masing-masing dimensi. Contohnya, data dari kolom agent akan dicocokkan dengan tabel dimensi Agent, kolom company dengan tabel Company, dan seterusnya, menggunakan Stream Lookup agar memperoleh key yang sesuai.

Setelah seluruh proses lookup selesai, dilakukan penambahan kolom booking\_id menggunakan komponen Add Sequence. Kolom ini berfungsi sebagai primary key unik untuk setiap baris pada fact table, memastikan tidak ada duplikasi dan memudahkan proses identifikasi transaksi atau pemesanan.

Selanjutnya, dibuat tiga kolom measure baru menggunakan komponen Calculator untuk menambah nilai-nilai yang bersifat kuantitatif dan penting untuk analisis, yaitu: total\_guest Merupakan jumlah total tamu, dihitung dari:  $total\_guest = adults\_total + children\_total + babies\_total$ , lalu total\_nights Menunjukkan total malam menginap dari gabungan hari biasa dan akhir pekan:  $total\_nights = stays\_in\_weekend\_nights + stays\_in\_week\_nights$ , dan revenue Merupakan estimasi pendapatan dari pemesanan:  $revenue = adr * total\_nights$  Di mana adr adalah Average Daily Rate atau rata-rata tarif per malam.

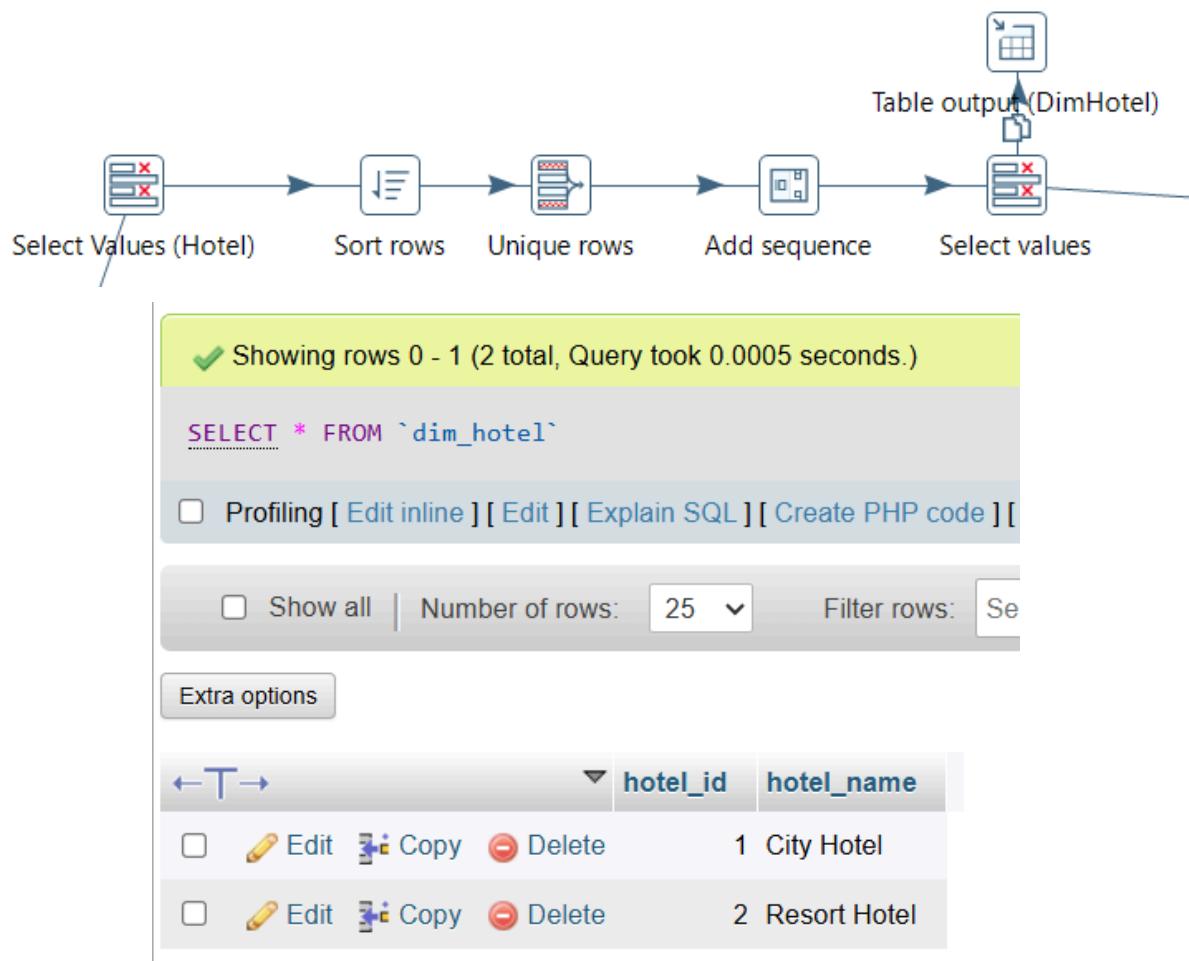
Struktur ini membentuk arsitektur Star Schema, di mana fact table berada di pusat dan dikelilingi oleh tabel-tabel dimensi — sehingga mendukung proses analisis data yang efisien dan terstruktur dalam sistem data warehouse.

## 3. Proses Load

Setelah seluruh proses transformasi selesai, tahap selanjutnya adalah proses load. Pada tahap ini, digunakan komponen Select Values terlebih dahulu untuk memilih kolom-kolom yang akan dimasukkan ke dalam fact table serta menyusun urutan kolom agar sesuai dengan rancangan star schema yang telah ditetapkan sebelumnya. Selain itu, dilakukan juga penyesuaian tipe data, salah satunya pada kolom reservation\_status\_date yang awalnya bertipe string, diubah menjadi tipe date agar sesuai dengan kebutuhan analisis waktu. Setelah semua kolom tersusun rapi dan tipe data telah disesuaikan, data hasil transformasi dimuat ke dalam database hotel\_dwbi menggunakan komponen Table Output, sehingga proses ETL terselesaikan dengan struktur data yang siap dianalisis.

## 5.2. Dimension Table

### a. dim\_hotel



Proses Extract, Transform, Load (ETL) pada hotel\_dwbi dalam Pentaho dilakukan dengan langkah-langkah berikut:

## **1. Proses Extract**

### a. CSV File Input

Mengambil data hotel dari sumber data berupa file hotel\_bookings.csv pada Pentaho.

## **2. Proses Transform**

Transformasi data dilakukan melalui beberapa tahapan dalam diagram ETL sebagai berikut:

### a. Select Values (Hotel)

Menyalin kolom “hotel” dari file input CSV dan mengganti nama kolom tersebut menjadi “hotel\_name” untuk keperluan analisis yang lebih jelas.

### b. Sort Rows

Menyortir data berdasarkan kolom hotel\_name untuk memudahkan proses identifikasi dan penanganan data duplikat secara konsisten.

### c. Unique Rows

Menghapus duplikasi berdasarkan kolom hotel\_name setelah proses penyortiran dilakukan. Langkah ini memastikan bahwa hanya data hotel yang unik yang akan dimasukkan ke dalam tabel dimensi.

### d. Add Sequence

Menambahkan kolom ID unik (biasanya berupa surrogate key auto-increment) untuk setiap baris data. Kolom ini berfungsi sebagai primary key dalam tabel dimensi — dalam hal ini, menghasilkan kolom hotel\_id untuk tabel dim\_hotel.

### e. Select Values

Digunakan untuk mengatur ulang urutan kolom agar data lebih rapi dan sesuai dengan struktur yang dibutuhkan untuk pemuatan ke dalam data warehouse.

## **3. Proses Load**

Proses load pada tabel dim\_hotel dilakukan dengan memasukkan data hotel yang telah dibersihkan dari duplikasi dan telah ditambahkan ID unik berupa hotel\_id. Data ini kemudian dimuat ke dalam tabel tujuan yang berada di dalam database hotel\_dwbi. Atribut yang dimasukkan hanya terdiri dari dua kolom, yaitu hotel\_id sebagai identitas unik setiap hotel dan hotel\_name sebagai nama hotel. Langkah ini bertujuan untuk membentuk tabel dimensi yang rapi, konsisten, dan siap digunakan dalam proses analisis data lebih lanjut di data warehouse.

- b. dim\_guest



Proses Extract, Transform, Load (ETL) pada hotel\_dwbi dalam Pentaho dilakukan dengan langkah-langkah berikut:

## 1. Proses Extract

- CSV File Input

Mengambil data pelanggan (adult, children, dan babies) dari sumber data berupa file hotel\_bookings.csv pada Pentaho.

## 2. Proses Transform

Transformasi data dilakukan melalui beberapa tahapan dalam diagram ETL sebagai berikut:

- Select Values (Guest)

Memilih tiga kolom utama dari file input CSV, yaitu adult, children, dan babies, tanpa melakukan proses penggantian nama (rename) terlebih dahulu.

- Sort Rows

Menyortir data berdasarkan ketiga kolom tersebut guna memudahkan identifikasi dan penanganan duplikasi secara konsisten.

c. Unique Rows

Menghapus baris-baris yang memiliki nilai duplikat berdasarkan kombinasi kolom adult, children, dan babies, sehingga hanya data kombinasi jumlah tamu yang unik yang akan dimuat ke dalam tabel dimensi.

d. Add Sequence

Menambahkan kolom ID unik (biasanya berupa surrogate key auto-increment) untuk setiap baris data. Kolom ini berfungsi sebagai primary key dalam tabel dimensi — dalam hal ini, menghasilkan kolom guest\_id untuk tabel dim\_guest.

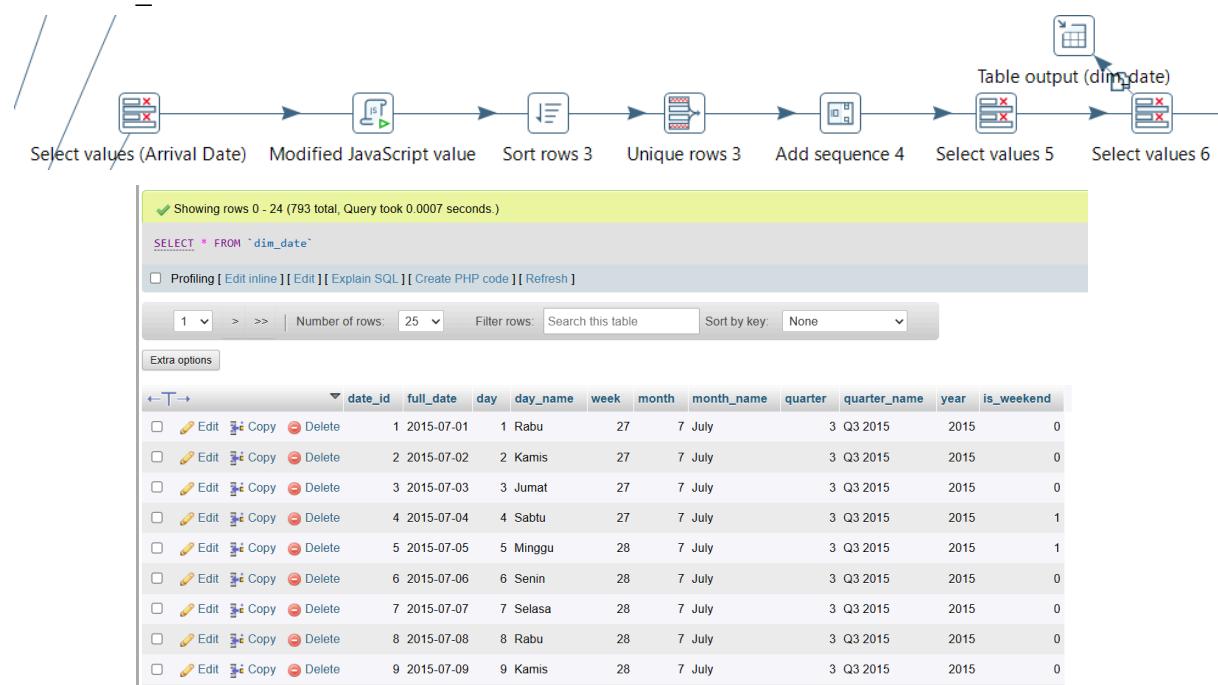
e. Select Values

Urutan kolom diatur ulang agar lebih rapi, sekaligus dilakukan penggantian nama kolom menjadi adult\_total, children\_total, dan babies\_total agar lebih representatif sebelum data dimuat ke dalam data warehouse.

### 3. Proses Load

Proses load pada tabel dim\_guest dilakukan dengan memasukkan data jumlah tamu yang telah dibersihkan dari duplikasi dan ditambahkan ID unik berupa guest\_id. Data ini kemudian dimuat ke dalam tabel tujuan yang berada di dalam database hotel\_dwbi. Atribut yang dimasukkan terdiri dari tiga kolom utama, yaitu adult\_total, children\_total, dan babies\_total, yang merepresentasikan jumlah tamu dewasa, anak-anak, dan bayi dalam setiap kombinasi unik. Penambahan guest\_id berfungsi sebagai identitas unik (surrogate key) untuk masing-masing kombinasi jumlah tamu. Langkah ini bertujuan untuk membentuk tabel dimensi dim\_guest yang rapi, konsisten, dan siap digunakan dalam proses analisis data lebih lanjut di data warehouse.

c. dim\_date



Proses Extract, Transform, Load (ETL) pada hotel\_dwbi dalam Pentaho dilakukan dengan langkah-langkah berikut:

## 1. Proses Extract

### a. CSV File Input

Data tanggal diekstrak dari file input CSV hotel booking dengan mengambil komponen tanggal kedatangan (arrival date) yang terdiri dari kolom arrival\_date\_year, arrival\_date\_month, arrival\_date\_week\_number, dan arrival\_date\_day\_of\_month.

## 2. Proses Transform

Transformasi data dilakukan melalui beberapa tahapan dalam diagram ETL sebagai berikut:

### a. Select Values (Guest)

Memilih empat kolom utama terkait tanggal dari file input CSV, yaitu arrival\_date\_year, arrival\_date\_month, arrival\_date\_week\_number, dan arrival\_date\_day\_of\_month, tanpa melakukan proses penggantian nama terlebih dahulu.

### b. Modified JavaScript Value

Transformasi ini melakukan beberapa operasi penting untuk membentuk dimensi tanggal yang komprehensif:

- Konversi nama bulan ke angka: Mengubah nama bulan (January hingga December) menjadi format numerik (01-12)
- Pembentukan full\_date: Menyusun format tanggal lengkap YYYY-MM-DD dengan menggabungkan tahun, bulan, dan hari
- Pembentukan date\_id: Membuat primary key unik dalam format YYYYMMDD (integer)
- Penentuan quarter: Mengklasifikasikan bulan ke dalam quarter (Q1-Q4) dan membentuk nama quarter (contoh: "Q1 2015")
- Perhitungan hari: Menggunakan tabel referensi dan algoritma Zeller sebagai fallback untuk menentukan nama hari (Senin-Minggu)
- Penentuan weekend: Menambahkan flag is\_weekend (1 untuk Sabtu-Minggu, 0 untuk hari lainnya)

### c. Sort Rows

Menyortir data berdasarkan komponen tanggal (arrival\_date\_year, month\_num, arrival\_date\_week\_number, arrival\_date\_day\_of\_month) untuk memudahkan identifikasi dan penanganan duplikasi.

### d. Unique Rows

Menghapus baris-baris duplikat berdasarkan kombinasi komponen tanggal, sehingga hanya data tanggal yang unik yang akan dimuat ke dalam tabel dimensi.

e. Add Sequence

Menambahkan kolom date\_id yang berfungsi sebagai surrogate key untuk tabel dim\_date.

f. Select Values

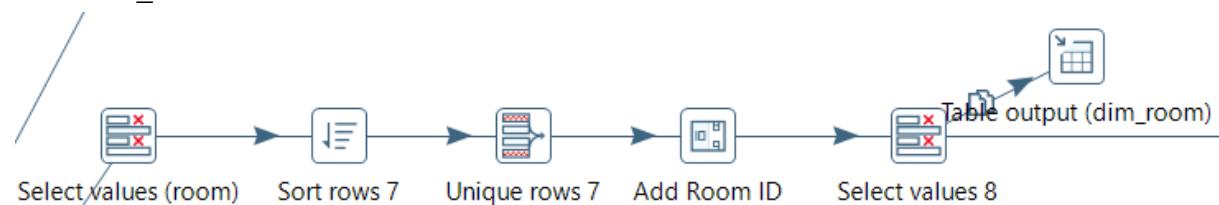
Mengatur ulang urutan kolom dan melakukan penggantian nama kolom sesuai dengan struktur tabel dimensi target:

- date\_id (Primary Key)
- full\_date (diubah dari full\_date\_str)
- day (diubah dari arrival\_date\_day\_of\_month)
- day\_name (nama hari dalam Bahasa Indonesia)
- week (diubah dari arrival\_date\_week\_number)
- month (nilai numerik bulan)
- month\_name (diubah dari arrival\_date\_month)
- quarter (nilai numerik quarter 1-4)
- quarter\_name (format "QX YYYY")
- year (diubah dari arrival\_date\_year)
- is\_weekend (flag untuk Sabtu dan Minggu)

### 3. Proses Load

Proses load pada tabel dim\_date dilakukan dengan memasukkan data tanggal yang telah ditransformasi ke dalam tabel tujuan dalam database hotel\_dwbi. Data ini mencakup keseluruhan dimensi waktu yang diperlukan untuk analisis seperti tanggal lengkap, hari, minggu, bulan, quarter, dan tahun serta indikator akhir pekan. Struktur dimensi waktu yang komprehensif ini memungkinkan analisis data berdasarkan berbagai granularitas waktu dan pola musiman, seperti tren pemesanan di akhir pekan versus hari kerja atau perbedaan perilaku pemesanan berdasarkan quarter tahun.

d. dim\_room



Showing rows 0 - 24 (75 total, Query took 0.0005 seconds.)

SELECT \* FROM `dim\_room`

Profiling [ Edit inline ] [ Edit ] [ Explain SQL ] [ Create PHP code ] [ Refresh ]

1 > >> | Show all | Number of rows: 25 Filter rows: Search this table Sort by key: None

Extra options

	room_id	reserved_room_category	assigned_room_category
<input type="checkbox"/>	1	A	A
<input type="checkbox"/>	2	A	B
<input type="checkbox"/>	3	A	C
<input type="checkbox"/>	4	A	D
<input type="checkbox"/>	5	A	E
<input type="checkbox"/>	6	A	F
<input type="checkbox"/>	7	A	G
<input type="checkbox"/>	8	A	H
<input type="checkbox"/>	9	A	I
<input type="checkbox"/>	10	A	K
<input type="checkbox"/>	11	B	A

Proses Extract, Transform, Load (ETL) pada hotel\_dwbi dalam Pentaho dilakukan dengan langkah-langkah berikut:

## 1. Proses Extract

### a. CSV File Input

Mengambil 2 data (reserved\_room\_type dan assigned\_room\_type) hotel dari sumber data berupa file hotel\_bookings.csv

## 2. Proses Transform

Transformasi data dilakukan melalui beberapa tahapan dalam diagram ETL sebagai berikut:

### a. Select Values (Hotel)

Pada tahap ini, dilakukan pengambilan 2 data, yaitu reserved\_room\_type dan assigned\_room\_type. Hal ini untuk memfokuskan data yang ingin dikelola

### b. Sort Rows

Menyortir data berdasarkan kolom reserved\_room\_type dan assigned\_room\_type dimana hal ini untuk memudahkan proses identifikasi dan penanganan data duplikat secara konsisten.

### c. Unique Rows

Menghapus duplikasi berdasarkan kolom reserved\_room\_type dan assigned\_room\_type setelah proses penyortiran dilakukan. Langkah ini memastikan bahwa hanya data room\_type yang unik yang akan dimasukkan ke dalam tabel dimensi.

### d. Add Sequence

Menambahkan kolom room\_id yang berfungsi sebagai surrogate key untuk tabel dim\_room.

e. Select Values

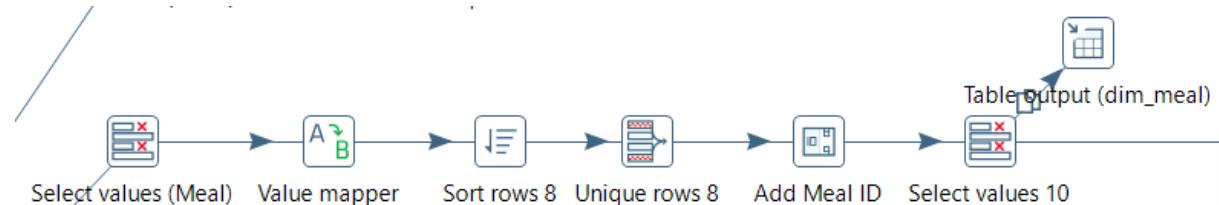
Mengatur ulang urutan kolom dan melakukan penggantian nama kolom sesuai dengan struktur tabel dimensi target:

- room\_id
- reserved\_room\_type
- assigned\_room\_type

### 3. Proses Load

Proses load pada tabel dim\_room dilakukan dengan memasukkan data kamar yang telah ditransformasi ke dalam tabel tujuan dalam database hotel\_dwbi. Data ini mencakup keseluruhan dimensi yang diperlukan untuk analisis seperti room\_id, reserved\_room\_category, dan assigned\_room\_category.

e. dim\_meal



meal_id	meal_code
1	BB
2	FB
3	HB
4	SC

Proses Extract, Transform, Load (ETL) pada hotel\_dwbi dalam Pentaho dilakukan dengan langkah-langkah berikut:

#### 1. Proses Extract

a. CSV File Input

Mengambil data meal dari sumber data berupa file hotel\_bookings.csv.

#### 2. Proses Transform

Transformasi data dilakukan melalui beberapa tahapan dalam diagram ETL sebagai berikut:

a. Select Values (Meal)

Menyalin kolom “meal” dari file input CSV untuk keperluan analisis yang lebih jelas.

b. Value Mapper

Mengganti data yang tidak valid pada kolom *meal*, yaitu nilai “Undefined”, dengan nilai “BB” yang merupakan modus (nilai yang paling sering muncul) pada kolom tersebut. Langkah ini dilakukan untuk memastikan konsistensi dan kualitas data sebelum digunakan dalam analisis lebih lanjut.

c. Sort Rows

Menyortir data berdasarkan kolom meal untuk memudahkan proses identifikasi dan penanganan data duplikat secara konsisten.

d. Unique Rows

Menghapus duplikasi berdasarkan kolom meal setelah proses penyortiran dilakukan. Langkah ini memastikan bahwa hanya data meal yang unik yang akan dimasukkan ke dalam tabel dimensi.

e. Add Sequence

Menambahkan kolom ID unik (biasanya berupa surrogate key auto-increment) untuk setiap baris data. Kolom ini berfungsi sebagai primary key dalam tabel dimensi — dalam hal ini, menghasilkan kolom meal\_id untuk tabel dim\_meal.

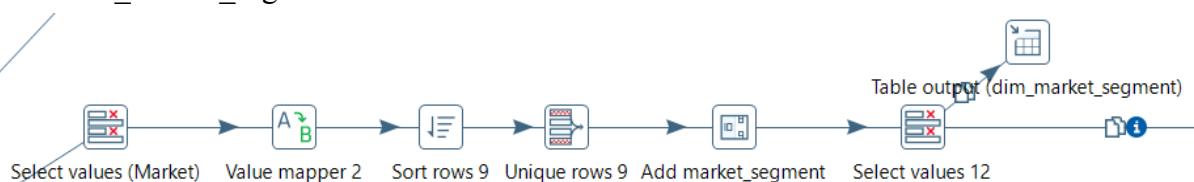
f. Select Values

Digunakan untuk mengatur ulang urutan kolom agar data lebih rapi dan sesuai dengan struktur yang dibutuhkan untuk pemuatan ke dalam data warehouse.

### 3. Proses Load

Proses load pada tabel dim\_meal dilakukan dengan memasukkan data meal yang telah dibersihkan dari duplikasi serta data yang tidak valid dan telah ditambahkan ID unik berupa meal\_id. Data ini kemudian dimuat ke dalam tabel tujuan yang berada di dalam database hotel\_dwbi. Atribut yang dimasukkan hanya terdiri dari dua kolom, yaitu meal\_id sebagai identitas unik setiap jenis makanan dan meal sebagai nama tipe makanan. Langkah ini bertujuan untuk membentuk tabel dimensi yang rapi, konsisten, dan siap digunakan dalam proses analisis data lebih lanjut di data warehouse.

f. dim\_market\_segment



		market_segment_id	market_segment_name	
<input type="checkbox"/>	<a href="#">Edit</a>	<a href="#">Copy</a>	<a href="#">Delete</a>	1 Aviation
<input type="checkbox"/>	<a href="#">Edit</a>	<a href="#">Copy</a>	<a href="#">Delete</a>	2 Complementary
<input type="checkbox"/>	<a href="#">Edit</a>	<a href="#">Copy</a>	<a href="#">Delete</a>	3 Corporate
<input type="checkbox"/>	<a href="#">Edit</a>	<a href="#">Copy</a>	<a href="#">Delete</a>	4 Direct
<input type="checkbox"/>	<a href="#">Edit</a>	<a href="#">Copy</a>	<a href="#">Delete</a>	5 Groups
<input type="checkbox"/>	<a href="#">Edit</a>	<a href="#">Copy</a>	<a href="#">Delete</a>	6 Offline TA/TO
<input type="checkbox"/>	<a href="#">Edit</a>	<a href="#">Copy</a>	<a href="#">Delete</a>	7 Online TA

Proses Extract, Transform, Load (ETL) pada hotel\_dwbi dalam Pentaho dilakukan dengan langkah-langkah berikut:

## 1. Proses Extract

### a. CSV File Input

Mengambil data market\_segment dari sumber data berupa file hotel\_bookings.csv.

## 2. Proses Transform

Transformasi data dilakukan melalui beberapa tahapan dalam diagram ETL sebagai berikut:

### a. Select Values (market\_segment)

Menyalin kolom “market\_segment” dari file input CSV untuk keperluan analisis yang lebih jelas.

### b. Value Mapper

Mengganti data yang tidak valid pada kolom *market\_segment*, yaitu nilai “Undefined”, dengan nilai “Online TA” yang merupakan modus (nilai yang paling sering muncul) pada kolom tersebut. Langkah ini dilakukan untuk memastikan konsistensi dan kualitas data sebelum digunakan dalam analisis lebih lanjut.

### c. Sort Rows

Menyortir data berdasarkan kolom market\_segment untuk memudahkan proses identifikasi dan penanganan data duplikat secara konsisten.

### d. Unique Rows

Menghapus duplikasi berdasarkan kolom market\_segment setelah proses penyortiran dilakukan. Langkah ini memastikan bahwa hanya data market\_segment yang unik yang akan dimasukkan ke dalam tabel dimensi.

### e. Add Sequence

Menambahkan kolom ID unik (biasanya berupa surrogate key auto-increment) untuk setiap baris data. Kolom ini berfungsi sebagai primary key dalam tabel dimensi — dalam hal ini, menghasilkan kolom market\_segment\_id untuk tabel dim\_market\_segment.

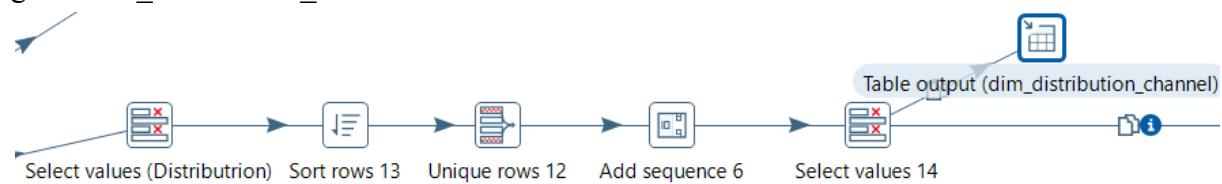
f. Select Values

Digunakan untuk mengatur ulang urutan kolom agar data lebih rapi dan sesuai dengan struktur yang dibutuhkan untuk pemuatan ke dalam data warehouse.

### 3. Proses Load

Proses load pada tabel dim\_market\_segment dilakukan dengan memasukkan data market segment yang telah dibersihkan dari duplikasi serta data yang tidak valid dan telah ditambahkan ID unik berupa market\_segment\_id. Data ini kemudian dimuat ke dalam tabel tujuan yang berada di dalam database hotel\_dwbi. Atribut yang dimasukkan hanya terdiri dari dua kolom, yaitu market\_segment\_id sebagai identitas unik setiap jenis segmen pasar dan market\_segment\_name sebagai nama tipe segmen pasar. Langkah ini bertujuan untuk membentuk tabel dimensi yang rapi, konsisten, dan siap digunakan dalam proses analisis data lebih lanjut di data warehouse.

g. dim\_distribution\_channel



	distribution_channel_id	distribution_channel_name
<input type="checkbox"/>	1	Corporate
<input type="checkbox"/>	2	Direct
<input type="checkbox"/>	3	GDS
<input type="checkbox"/>	4	TA/TO

Proses Extract, Transform, Load (ETL) pada hotel\_dwbi dalam Pentaho dilakukan dengan langkah-langkah berikut:

#### 1. Proses Extract

a. CSV File Input

Mengambil data distribution\_channel dari sumber data berupa file hotel\_bookings.csv.

## 2. Proses Transform

Transformasi data dilakukan melalui beberapa tahapan dalam diagram ETL sebagai berikut:

a. Select Values (Distribution\_Channel)

Menyalin kolom “distribution\_channel” dari file input CSV untuk keperluan analisis yang lebih jelas.

b. Value Mapper

Mengganti data yang tidak valid pada kolom distribution\_channel, yaitu nilai “Undefined”, dengan nilai “TA/TO” yang merupakan modus (nilai yang paling sering muncul) pada kolom tersebut. Langkah ini dilakukan untuk memastikan konsistensi dan kualitas data sebelum digunakan dalam analisis lebih lanjut.

c. Sort Rows

Menyortir data berdasarkan kolom distribution\_channel untuk memudahkan proses identifikasi dan penanganan data duplikat secara konsisten.

d. Unique Rows

Menghapus duplikasi berdasarkan kolom distribution\_channel setelah proses penyortiran dilakukan. Langkah ini memastikan bahwa hanya data distribution\_channel yang unik yang akan dimasukkan ke dalam tabel dimensi.

e. Add Sequence

Menambahkan kolom ID unik (biasanya berupa surrogate key auto-increment) untuk setiap baris data. Kolom ini berfungsi sebagai primary key dalam tabel dimensi, dimana menghasilkan kolom distribution\_channel\_id untuk tabel dim\_distribution\_channel.

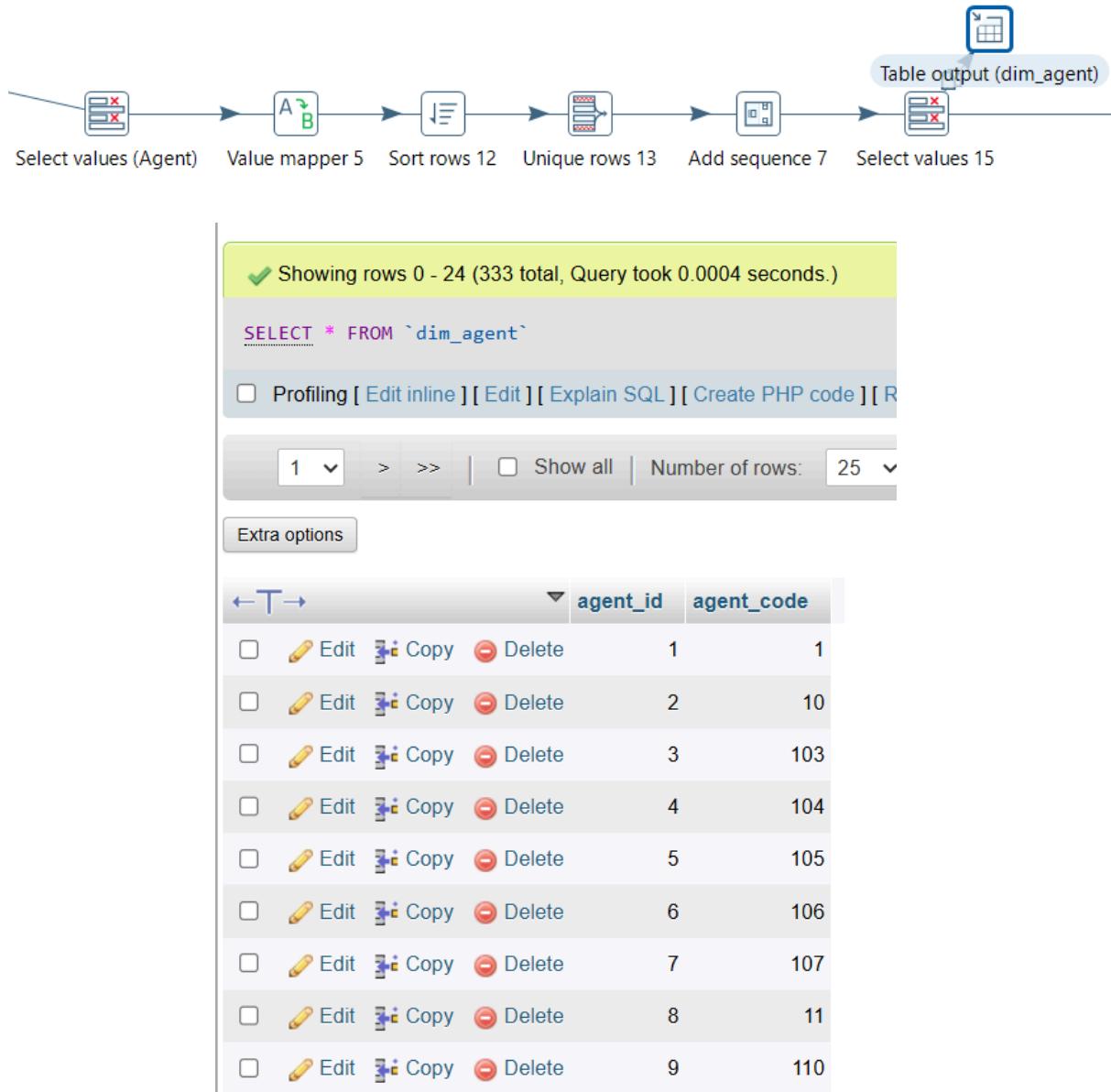
f. Select Values

Digunakan untuk mengatur ulang urutan kolom agar data lebih rapi dan sesuai dengan struktur yang dibutuhkan untuk pemuatan ke dalam data warehouse.

## 3. Proses Load

Proses load pada tabel dim\_meal dilakukan dengan memasukkan data meal yang telah dibersihkan dari duplikasi serta data yang tidak valid dan telah ditambahkan ID unik berupa distribution\_channel\_id. Data ini kemudian dimuat ke dalam tabel tujuan yang berada di dalam database. Atribut yang dimasukkan hanya terdiri dari dua kolom, yaitu distribution\_channel\_id sebagai identitas unik setiap distribusinya dan distribution\_channel\_name untuk nama tiap channel distribusi. Langkah ini bertujuan untuk membentuk tabel dimensi yang rapi, konsisten, dan siap digunakan dalam proses analisis data lebih lanjut di data warehouse.

h. dim\_agent



Proses Extract, Transform, Load (ETL) pada hotel\_dwbi dalam Pentaho dilakukan dengan langkah-langkah berikut:

## 1. Proses Extract

### a. CSV File Input

Mengambil data dari dim\_agent dari sumber data berupa file hotel\_bookings.csv.

## 2. Proses Transform

Transformasi data dilakukan melalui beberapa tahapan dalam diagram ETL sebagai berikut:

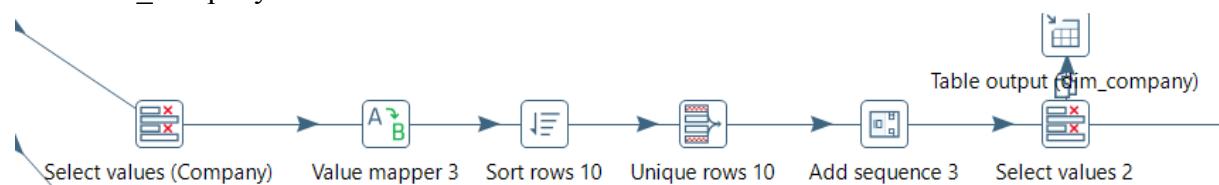
### a. Select values (Agent)

- Mengambil data “agent” dari file input CSV untuk keperluan analisis yang lebih jelas.
- Value mapper  
Mengganti data NULL pada kolom agent dengan nilai “0” yang mengindikasikan bahwa transaksi booking tersebut tidak dilakukan oleh sebuah agent. Langkah ini dilakukan untuk memastikan konsistensi dan kualitas data sebelum digunakan dalam analisis lebih lanjut.
  - Sort rows  
Menyortir data berdasarkan kolom agent untuk memudahkan proses identifikasi dan penanganan data duplikat secara konsisten.
  - Unique rows  
Menghapus duplikasi berdasarkan kolom agent setelah proses penyortiran dilakukan. Langkah ini memastikan bahwa hanya data agent yang unik yang akan dimasukkan ke dalam tabel dimensi.
  - Add sequence  
Menambahkan kolom ID unik (biasanya berupa surrogate key auto-increment) untuk setiap baris data. Kolom ini berfungsi sebagai primary key dalam tabel dimensi, dimana menghasilkan kolom agent\_id untuk tabel dim\_agent.
  - Select values  
Digunakan untuk mengatur ulang urutan kolom agar data lebih rapi dan sesuai dengan struktur yang dibutuhkan untuk pemuatan ke dalam data warehouse. Dimana ada perubahan seperti adanya value tambahan yakni “agent\_id” dan perubahan nama yang awalnya “agent” menjadi “agent\_code”.

### 3. Proses Load

Proses load pada tabel dim\_agent dilakukan dengan memasukkan data market segment yang telah dibersihkan dari duplikasi serta data yang tidak valid dan telah ditambahkan ID unik berupa agent\_id. Data ini kemudian dimuat ke dalam tabel tujuan yang berada di dalam database hotel\_dwbi. Atribut yang dimasukkan hanya terdiri dari dua kolom, yaitu agent\_id sebagai identitas unik setiap jenis agent dan agent berubah menjadi agent\_code sebagai kode dari agent. Langkah ini bertujuan untuk membentuk tabel dimensi yang rapi, konsisten, dan siap digunakan dalam proses analisis data lebih lanjut di data warehouse.

#### i. dim\_company



The screenshot shows a MySQL query results page. At the top, a green bar displays the message "Showing rows 0 - 24 (353 total, Query took 0.0004 seconds.)". Below this is the SQL query: "SELECT \* FROM `dim\_company`". A toolbar below the query includes options for Profiling, Edit inline, Edit, Explain SQL, Create PHP code, and Refresh. Navigation controls (1, >, >>) and a search/filter bar are also present. The main area displays a table with columns company\_id and company\_code. The data is as follows:

	company_id	company_code
<input type="checkbox"/>	1	0
<input type="checkbox"/>	2	10
<input type="checkbox"/>	3	100
<input type="checkbox"/>	4	101
<input type="checkbox"/>	5	102
<input type="checkbox"/>	6	103
<input type="checkbox"/>	7	104
<input type="checkbox"/>	8	105
<input type="checkbox"/>	9	106
<input type="checkbox"/>	10	107

Proses Extract, Transform, Load (ETL) pada hotel\_dwbi dalam Pentaho dilakukan dengan langkah-langkah berikut:

## 1. Proses Extract

### a. CSV File Input

Mengambil data dari dim\_company dari sumber data berupa file hotel\_bookings.csv.

## 2. Proses Transform

Transformasi data dilakukan melalui beberapa tahapan dalam diagram ETL sebagai berikut:

### a. Select Values (Company)

Menyalin kolom “company” dari file input CSV untuk keperluan analisis yang lebih jelas.

### b. Value Mapper

Mengganti data NULL pada kolom company dengan nilai “0” yang mengindikasikan bahwa transaksi booking tersebut tidak diwakilkan oleh sebuah perusahaan. Langkah ini dilakukan untuk memastikan konsistensi dan kualitas data sebelum digunakan dalam analisis lebih lanjut.

### c. Sort Rows

Menyortir data berdasarkan kolom company untuk memudahkan proses identifikasi dan penanganan data duplikat secara konsisten.

d. Unique Rows

Menghapus duplikasi berdasarkan kolom company setelah proses penyortiran dilakukan. Langkah ini memastikan bahwa hanya data company yang unik yang akan dimasukkan ke dalam tabel dimensi.

e. Add Sequence

Menambahkan kolom ID unik (biasanya berupa surrogate key auto-increment) untuk setiap baris data. Kolom ini berfungsi sebagai primary key dalam tabel dimensi, dimana menghasilkan kolom company\_id untuk tabel dim\_company.

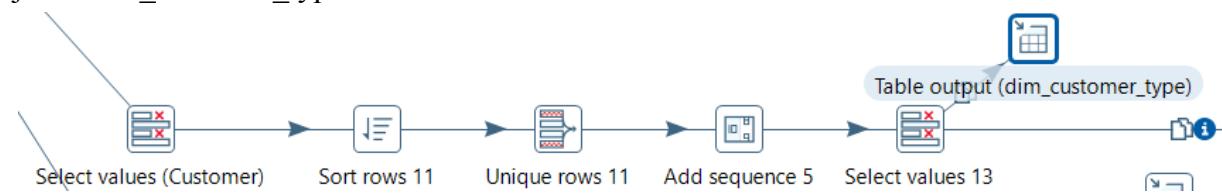
f. Select Values

Digunakan untuk mengatur ulang urutan kolom agar data lebih rapi dan sesuai dengan struktur yang dibutuhkan untuk pemuatan ke dalam data warehouse.

### 3. Proses Load

Proses load pada tabel dim\_company dilakukan dengan memasukkan data company yang telah dibersihkan dari duplikasi serta nilai NULL dan telah ditambahkan ID unik berupa company\_id. Data ini kemudian dimuat ke dalam tabel tujuan yang berada di dalam database. Atribut yang dimasukkan hanya terdiri dari dua kolom, yaitu company\_id sebagai identitas unik setiap perusahaan dan company\_code sebagai nama jenis-jenis perusahaan dengan nilai 0 berarti individu. Langkah ini bertujuan untuk membentuk tabel dimensi yang rapi, konsisten, dan siap digunakan dalam proses analisis data lebih lanjut di data warehouse.

j. dim\_customer\_type



<span style="color: green;">✓</span> Showing rows 0 - 3 (4 total, Query took 0.0004 seconds.)																
<pre>SELECT * FROM `dim_customer_type`</pre>																
<input type="checkbox"/> Profiling [ Edit inline ] [ Edit ] [ Explain SQL ] [ Create PHP code ] [ Refresh ]																
<input type="checkbox"/> Show all   Number of rows: <select>25</select> Filter rows: <input type="text" value="Search this table"/>																
<input type="button" value="Extra options"/>																
<table border="1"> <thead> <tr> <th></th> <th>customer_type_id</th> <th>customer_category</th> </tr> </thead> <tbody> <tr> <td><input type="checkbox"/></td><td></td><td>1 Contract</td></tr> <tr> <td><input type="checkbox"/></td><td></td><td>2 Group</td></tr> <tr> <td><input type="checkbox"/></td><td></td><td>3 Transient</td></tr> <tr> <td><input type="checkbox"/></td><td></td><td>4 Transient-Party</td></tr> </tbody> </table>			customer_type_id	customer_category	<input type="checkbox"/>		1 Contract	<input type="checkbox"/>		2 Group	<input type="checkbox"/>		3 Transient	<input type="checkbox"/>		4 Transient-Party
	customer_type_id	customer_category														
<input type="checkbox"/>		1 Contract														
<input type="checkbox"/>		2 Group														
<input type="checkbox"/>		3 Transient														
<input type="checkbox"/>		4 Transient-Party														

Proses Extract, Transform, Load (ETL) pada hotel\_dwbi dalam Pentaho dilakukan dengan langkah-langkah berikut:

## 1. Proses Extract

- a. CSV File Input

Mengambil data pelanggan yaitu “customer\_type” dari sumber data berupa file hotel\_bookings.csv pada Pentaho

## 2. Proses Transform

Transformasi data dilakukan melalui beberapa tahapan dalam diagram ETL sebagai berikut:

- a. Select values (Customer\_Type)

Memilih kolom “customer\_type” dari file input untuk keperluan analisis yang lebih jelas.

- b. Sort rows 13

Menyortir data berdasarkan kolom customer\_type untuk memudahkan proses identifikasi dan penanganan data duplikat secara konsisten.

- c. Unique rows 12

Menghapus duplikasi berdasarkan kolom customer\_type setelah proses penyortiran dilakukan. Langkah ini memastikan bahwa hanya data customer\_type yang unik yang akan dimasukkan ke dalam tabel dimensi.

- d. Add sequence 6

Menambahkan kolom ID unik yaitu customer\_type\_id (biasanya berupa surrogate key auto-increment) untuk setiap baris data. Kolom ini berfungsi sebagai primary key dalam tabel dimensi. dalam hal ini, menghasilkan kolom customer\_type\_id untuk tabel dim\_customer\_type.

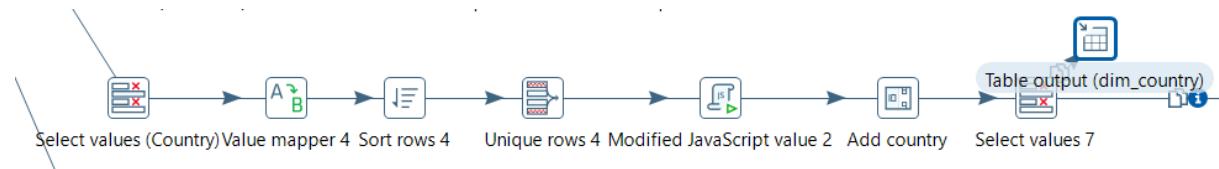
e. Select values 14

Digunakan untuk mengatur ulang urutan kolom customer\_type agar data lebih rapi dan sesuai dengan struktur yang dibutuhkan untuk pemuatan ke dalam data warehouse.

### 3. Proses Load

Proses load pada tabel dim\_customer\_type dilakukan dengan memasukkan data customer type yang telah dibersihkan dari duplikasi dan telah ditambahkan ID unik berupa customer\_type\_id. Data ini kemudian dimuat ke dalam tabel tujuan yang berada di dalam database hotel\_dwbi. Atribut yang dimasukkan hanya terdiri dari dua kolom, yaitu customer\_type\_id sebagai identitas unik setiap tipe costumer dan customer\_type sebagai nama kategori pelanggan. Langkah ini bertujuan untuk membentuk tabel dimensi yang rapi, konsisten, dan siap digunakan dalam proses analisis data lebih lanjut di data warehouse.

k. dim\_country



The screenshot shows a database query results window with the following details:

- Header: "Showing rows 0 - 24 (176 total, Query took 0.0006 seconds.)"
- SQL Query: "SELECT \* FROM `dim\_country`"
- Buttons: Profiling [ Edit inline ], Edit, Explain SQL, Create PHP code, Refresh.
- Table Headers: country\_id, country\_code, country\_name.
- Data Rows (partial):

country_id	country_code	country_name
1	ABW	Aruba
2	AGO	Angola
3	AIA	Anguilla
4	ALB	Albania
5	AND	Andorra
6	ARE	United Arab Emirates
7	ARG	Argentina
8	ARM	Armenia
9	ASM	American Samoa
10	ATA	Antarctica
11	ATF	Unknown (ATF)

Proses Extract, Transform, Load (ETL) pada hotel\_dwbi dalam Pentaho dilakukan dengan langkah-langkah berikut:

## 1. Proses Extract

### a. CSV File Input

Mengambil data negara di dim\_country dari sumber data berupa file hotel\_bookings.csv.

## 2. Proses Transform

Transformasi data dilakukan melalui beberapa tahapan dalam diagram ETL sebagai berikut:

- Select Values (Distribution\_Channel)

Menyalin kolom "country" dari file input CSV untuk keperluan analisis yang lebih jelas.

- Value Mapper

Mengganti data "CN" dan "NULL" pada kolom company dengan nilai "CHN" yang berarti negara China dan nilai NULL dengan nilai modus yaitu PRT (Portugal). Langkah ini dilakukan untuk memastikan konsistensi dan kualitas data sebelum digunakan dalam analisis lebih lanjut.

- Sort Rows

Menyortir data berdasarkan kolom country untuk memudahkan proses identifikasi dan penanganan data duplikat secara konsisten.

- Unique Rows

Menghapus duplikasi berdasarkan kolom country setelah proses penyortiran dilakukan. Langkah ini memastikan bahwa hanya data country yang unik yang akan dimasukkan ke dalam tabel dimensi.

- Modified JavaScript Value

A. Transformasi ini melakukan pemetaan kode negara menjadi nama negara lengkap yang lebih mudah dibaca dan digunakan untuk analisis. Transformasi ini mencakup langkah-langkah berikut:

B. Pemetaan Kode Negara: Menggunakan objek countryMap, setiap kode negara tiga huruf (seperti "IDN", "USA", atau "FRA") dipetakan ke nama negara lengkap (seperti "Indonesia", "United States", atau "France").

C. Validasi Kode: Jika kode negara tidak ditemukan dalam daftar countryMap, maka nilai default "Unknown (kode\_negara)" akan diberikan sebagai penanda bahwa kode tersebut tidak dikenali.

D. Logging Kesalahan: Jika kode tidak dikenal, sistem akan mencatat pesan peringatan (log level: WARNING) untuk memudahkan debugging atau validasi data selanjutnya.

- Add Sequence

Menambahkan kolom ID unik (biasanya berupa surrogate key auto-increment) untuk setiap baris data. Kolom ini berfungsi sebagai primary key dalam tabel dimensi, dimana menghasilkan kolom country\_id untuk tabel dim\_country.

- Select Values

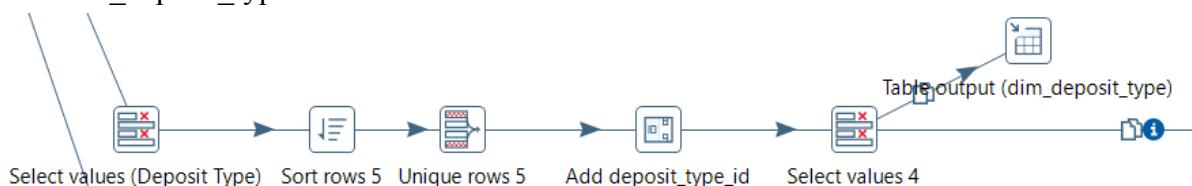
Digunakan untuk mengatur ulang urutan kolom agar data lebih rapi dan sesuai dengan struktur yang dibutuhkan untuk pemuatan ke dalam data warehouse.

Dimana tahap ini juga, data yang awalnya “country” di rename menjadi “country\_code”

### 3. Proses Load

Proses load pada tabel dim\_country dilakukan dengan memasukkan data country yang telah dibersihkan dari duplikasi serta nilai NULL dan tidak valid, lalu telah ditambahkan ID unik berupa country\_id. Data ini kemudian dimuat ke dalam tabel tujuan yang berada di dalam database. Atribut yang dimasukkan hanya terdiri dari tiga kolom, yaitu country\_id sebagai identitas unik setiap negara, country\_code yaitu kode unik ISO tiap negara, dan country\_name sebagai nama lengkap negara hasil translasi kode negara ISO. Langkah ini bertujuan untuk membentuk tabel dimensi yang rapi, konsisten, dan siap digunakan dalam proses analisis data lebih lanjut di data warehouse.

#### 1. dim\_deposit\_type



	deposit_type_id	deposit_category
<input type="checkbox"/> <a href="#">Edit</a> <a href="#">Copy</a> <a href="#">Delete</a>	1	No Deposit
<input type="checkbox"/> <a href="#">Edit</a> <a href="#">Copy</a> <a href="#">Delete</a>	2	Non Refund
<input type="checkbox"/> <a href="#">Edit</a> <a href="#">Copy</a> <a href="#">Delete</a>	3	Refundable

Proses Extract, Transform, Load (ETL) pada hotel\_dwbi dalam Pentaho dilakukan dengan langkah-langkah berikut:

#### 1. Proses Extract

##### a. CSV File Input

Mengambil data deposit yaitu “deposit\_type” dari sumber data berupa file hotel\_bookings.csv pada Pentaho

## 2. Proses Transform

Transformasi data dilakukan melalui beberapa tahapan dalam diagram ETL sebagai berikut:

- Select values (Deposit Type)

Proses ini mengambil hanya kolom deposit type dari data sumber. Tujuannya adalah memisahkan atribut yang akan dimasukkan ke dalam tabel dimensi dim\_deposit\_type.

- Sort rows

Menyortir data berdasarkan kolom deposit\_type untuk memudahkan proses identifikasi dan penanganan data duplikat secara konsisten.

- Unique rows

Menghapus duplikasi berdasarkan kolom deposit\_type setelah proses penyortiran dilakukan. Langkah ini memastikan bahwa hanya data hotel yang unik yang akan dimasukkan ke dalam tabel dimensi.

- Add sequence\_deposit\_type

Menambahkan kolom ID unik (surrogate key) yaitu deposit\_type\_id untuk setiap baris deposit type. Biasanya ID ini bersifat auto-increment atau menggunakan urutan tertentu agar bisa dijadikan primary key.

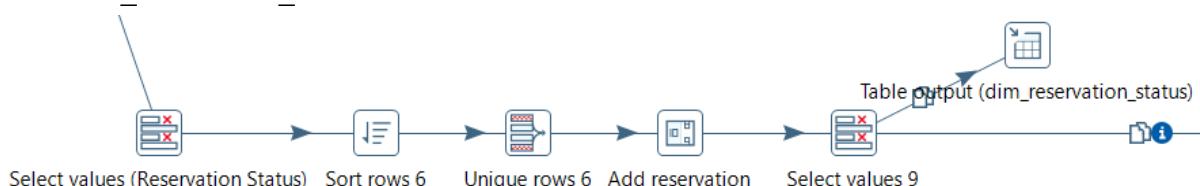
- Select values

Digunakan untuk mengatur ulang urutan kolom deposit\_type agar data lebih rapi dan sesuai dengan struktur yang dibutuhkan untuk pemuatan ke dalam data warehouse. Memilih kolom deposit\_type\_id dan deposit\_type yang direname menjadi deposit\_category saja untuk dimasukkan ke dalam tabel target. Langkah ini memastikan hanya kolom relevan yang ditulis ke dalam tabel dimensi.

## 3. Proses Load

Data hasil transformasi dimasukkan ke dalam tabel dimensi dim\_deposit\_type di database hotel\_dwbi, yang memiliki dua kolom deposit\_type\_id (primary key) dan deposit\_category. Langkah ini bertujuan untuk membentuk tabel dimensi yang rapi, konsisten, dan siap digunakan dalam proses analisis data lebih lanjut di data warehouse.

- dim\_reservation\_status



Showing rows 0 - 2 (3 total, Query took 0.0004 seconds.)

```
SELECT * FROM `dim_reservation_status`
```

Profiling | [Edit inline](#) | [Edit](#) | [Explain SQL](#) | [Create PHP code](#) | [Refresh](#)

Show all | Number of rows: 25 Filter rows: Search this table

[Extra options](#)

	reservation_status_id	reservation_stats
<input type="checkbox"/>	1	Canceled
<input type="checkbox"/>	2	Check-Out
<input type="checkbox"/>	3	No-Show

Proses Extract, Transform, Load (ETL) pada dim\_hotel dalam Pentaho dilakukan dengan langkah-langkah berikut:

## 1. Proses Extract

### a. CSV File Input

Mengambil data “reservation\_status” dari sumber data berupa file hotel\_bookings.csv pada Pentaho

## 2. Proses Transform

Transformasi data dilakukan melalui beberapa tahapan dalam diagram ETL sebagai berikut:

### a. Select values (Reservation Status)

Memilih kolom “Reservation\_status” dari file input.

### b. Sort rows

Menyortir data berdasarkan kolom reservation\_status untuk memudahkan proses identifikasi dan penanganan data duplikat secara konsisten.

### c. Unique rows

Menghapus duplikasi berdasarkan kolom reservation\_status setelah proses penyortiran dilakukan. Langkah ini memastikan bahwa hanya data status reservasi yang unik yang akan dimasukkan ke dalam tabel dimensi.

### d. Add Sequence

Menambahkan kolom ID unik (biasanya berupa surrogate key auto-increment). Kolom ini berfungsi sebagai primary key dalam tabel dimensi dalam hal ini, menghasilkan kolom reservation\_status\_id untuk tabel dim\_reservation\_status.

### e. Select Values

Memilih hanya dua kolom penting yaitu reserved\_status\_id dan reservation\_status untuk dimasukkan ke tabel tujuan. Digunakan untuk mengatur ulang urutan kolom agar data lebih rapi dan sesuai dengan struktur yang dibutuhkan untuk pemuatan ke dalam data warehouse.

### **3. Proses Load**

Proses load pada tabel dim\_reservation\_status dilakukan dengan memasukkan data reservation\_status yang telah dibersihkan dari duplikasi dan telah ditambahkan ID unik berupa reservation\_status\_id. Data ini kemudian dimuat ke dalam tabel tujuan yang berada di dalam database hotel\_dwbi. Atribut yang dimasukkan hanya terdiri dari dua kolom, yaitu reservation\_status\_id sebagai identitas unik setiap status reservasi dan reservation\_status sebagai status dari tiap reservasi. Langkah ini bertujuan untuk membentuk tabel dimensi yang rapi, konsisten, dan siap digunakan dalam proses analisis data lebih lanjut di data warehouse.

# BAB VI IMPLEMENTASI DATA MINING

## 6.1. Klasifikasi (CatBoost Classifier)

Klasifikasi adalah teknik machine learning yang memprediksi kategori atau kelas target berdasarkan fitur-fitur input. Dalam penelitian ini, kami menggunakan algoritma CatBoost Classifier untuk memprediksi apakah suatu pemesanan hotel akan dibatalkan atau tidak (kolom `is_canceled`). CatBoost dipilih karena keunggulannya dalam menangani data kategorikal secara otomatis tanpa perlu encoding manual, kemampuannya mengatasi missing values, performa yang superior dibandingkan algoritma tradisional, serta resistensi terhadap overfitting berkat implementasi gradient boosting dengan teknik permutasi yang inovatif. Model ini sangat cocok untuk dataset hotel booking yang memiliki kombinasi fitur numerik dan kategorikal, sehingga dapat mengoptimalkan prediksi pembatalan reservasi dan membantu manajemen hotel mengantisipasi ketersediaan kamar secara lebih akurat.

### 6.1.1. Data Cleaning & Preparation

#### a. Outlier Handling

```
numerical_features = df_train.select_dtypes(include=['int64', 'float64']).columns.tolist()
✓ 0.0s

def remove_outliers(df_train, column):
    Q1 = df_train[column].quantile(0.25)
    Q3 = df_train[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return df_train[(df_train[column] >= lower_bound) & (df_train[column] <= upper_bound)]

for col in numerical_features:
    data_outlier = remove_outliers(df_train, col)
✓ 0.5s

print("Total row before remove outlier:")
print(df_train.shape)
print("\nTotal row after remove outlier:")
print(data_outlier.shape)
✓ 0.0s

Total row before remove outlier:
(119390, 34)

Total row after remove outlier:
(116513, 34)
```

Penanganan outlier dilakukan dengan metode robust IQR (Interquartile Range) untuk meningkatkan kualitas dan keandalan model. Implementasinya meliputi identifikasi fitur numerik menggunakan `df_train.select_dtypes(include=['int64', 'float64'])`, dilanjutkan dengan fungsi `remove_outliers()` yang menghitung batas bawah ( $Q1 - 1.5 \times IQR$ ) dan batas atas ( $Q3 + 1.5 \times IQR$ ) untuk setiap kolom numerik. Data yang berada di luar batas tersebut dihapus, menghasilkan dataset yang lebih bersih dengan berkurangnya jumlah baris dari 119.390 menjadi 116.513 (eliminasi sekitar 2.4% data). Pendekatan ini memastikan anomali-anomali ekstrim tidak mempengaruhi

performa model klasifikasi dan clustering, sehingga menghasilkan analisis yang lebih akurat dan representatif terhadap pola umum data pemesanan hotel.

## b. Null Values Handling

```
# Imputasi dan fitur baru untuk 'company'
data_outlier['company'] = data_outlier['company'].fillna(0)
data_outlier['company'] = data_outlier['company'].astype(int)
data_outlier['is_company'] = data_outlier['company'].apply(lambda x: 0 if x == 0 else 1)
data_outlier = data_outlier.drop(columns=['company'])
✓ 0.0s

# Imputasi dan fitur baru untuk 'agent'
data_outlier['agent'] = data_outlier['agent'].fillna(0)
data_outlier['agent'] = data_outlier['agent'].astype(int)
data_outlier['is_agent'] = data_outlier['agent'].apply(lambda x: 0 if x == 0 else 1)
data_outlier = data_outlier.drop(columns=['agent'])
✓ 0.0s

# Imputasi 'country' dengan modus
modus_country = data_outlier['country'].mode()[0]
data_outlier['country'] = data_outlier['country'].fillna(modus_country)
✓ 0.0s

# Imputasi 'children' dengan 0 dan ubah ke int
data_outlier['children'] = data_outlier['children'].fillna(0)
data_outlier['children'] = data_outlier['children'].astype(int)
✓ 0.0s
```

Penanganan nilai null dilakukan dengan pemahaman kontekstual terhadap makna bisnis setiap kolom. Untuk kolom 'company' dan 'agent' yang memiliki persentase missing values tinggi, nilai null menandakan bahwa pemesanan tidak dilakukan melalui perusahaan atau agen perjalanan (reservasi langsung oleh tamu), sehingga diimputasi dengan nilai 0 dan dikonversi ke tipe integer. Pendekatan ini diperkuat dengan pembuatan fitur biner ('is\_company' dan 'is\_agent') yang bernilai 0 jika pemesanan tanpa perantara dan 1 jika melalui perusahaan/agen, memberikan model informasi eksplisit tentang jalur pemesanan. Kolom 'country' diimputasi dengan nilai modus untuk mempertahankan distribusi geografis tamu, sedangkan 'children' yang hanya memiliki sedikit nilai null diisi dengan 0 mengasumsikan tidak ada anak dalam pemesanan tersebut. Strategi ini mempertahankan integritas semantik data sesuai realitas bisnis perhotelan.

## c. Handling Invalid Values

```

cols_with_undefined = [
    'market_segment',
    'meal',
    'distribution_channel'
]

for col in cols_with_undefined:
    modus = data_outlier.loc[data_outlier[col] != 'Undefined', col].mode()[0]
    data_outlier[col] = data_outlier[col].replace('Undefined', modus)

✓ 0.0s

```

Dalam penanganan nilai yang tidak valid, pendekatan yang diambil berfokus pada kolom kategorikal yang memiliki entri "Undefined". Tiga kolom utama yang diidentifikasi memiliki masalah tersebut adalah 'market\_segment', 'meal', dan 'distribution\_channel' - ketiganya merupakan fitur penting yang menggambarkan jalur pemasaran dan preferensi tamu. Untuk setiap kolom ini, nilai "Undefined" diganti dengan modus (nilai yang paling sering muncul) dari masing-masing kolom, sehingga mempertahankan distribusi data yang representatif. Pendekatan ini lebih baik daripada menghapus baris yang mengandung nilai tidak valid, karena memungkinkan retensi data yang lebih tinggi untuk analisis. Proses standardisasi ini memastikan konsistensi dalam kategori yang digunakan, yang sangat penting untuk akurasi model klasifikasi CatBoost yang mengandalkan data kategorikal yang bersih untuk performa optimal.

#### d. Handling Identical Features with Target

```

drop_cols = ['reservation_status', 'reservation_status_date']
data_outlier = data_outlier.drop(columns=drop_cols)

✓ 0.0s

```

Untuk mencegah data leakage, dilakukan penghapusan fitur 'reservation\_status' dan 'reservation\_status\_date' yang memiliki hubungan langsung dan identik dengan target prediksi 'is\_canceled'. Hal ini sangat krusial karena kedua kolom tersebut mengandung informasi yang secara langsung mengekspos status pembatalan - 'reservation\_status' bahkan secara eksplisit berisi nilai "Canceled" untuk pemesanan yang dibatalkan. Penggunaan fitur ini sebelumnya menghasilkan model dengan akurasi mencapai 1.0 (100%), yang jelas menunjukkan terjadinya data leakage dan model tidak belajar pola yang sebenarnya. Dengan menghapus kolom-kolom ini, model dipaksa untuk mempelajari pola dari fitur-fitur lain yang lebih representatif terhadap kondisi sebelum keputusan pembatalan dibuat, sehingga menghasilkan model yang lebih realistik dan dapat dipercaya untuk implementasi di dunia nyata.

#### e. Discretization (Data Binning)

```

# Binning lead_time
bins = [0, 7, 30, 60, 1000]
labels = ['Sangat Pendek', 'Pendek', 'Menengah', 'Panjang']
data_outlier['lead_time_bin'] = pd.cut(data_outlier['lead_time'], bins=bins, labels=labels, right=True, include_lowest=True)

# Tampilkan value count hasil binning
data_outlier['lead_time_bin'].value_counts()

```

lead_time_bin	count
Panjang	62172
Sangat Pendek	19359
Pendek	18427
Menengah	16555
Name: count, dtype: int64	

Diskretisasi dilakukan pada fitur 'lead\_time' (waktu antara pemesanan dan check-in) untuk mentransformasikan variabel kontinu yang memiliki rentang nilai luas (0-737 hari) menjadi kategori yang lebih bermakna. Proses binning ini membagi nilai 'lead\_time' ke dalam empat kategori yang diberi label intuitif: 'Sangat Pendek' (0-7 hari), 'Pendek' (8-30 hari), 'Menengah' (31-60 hari), dan 'Panjang' (>60 hari). Hasil pengelompokan menunjukkan distribusi yang tidak seimbang dengan mayoritas pemesanan (62.172 entri atau sekitar 53%) termasuk kategori 'Panjang', diikuti oleh kategori 'Sangat Pendek' (19.359 entri), 'Pendek' (18.427 entri), dan 'Menengah' (16.555 entri). Teknik ini tidak hanya membantu mengurangi kompleksitas data dan meningkatkan interpretabilitas, tetapi juga memberikan perspektif segmentasi penting terkait perilaku pemesanan pelanggan, yang sangat bermanfaat untuk analisis strategi pemasaran berdasarkan jendela waktu pemesanan.

## f. Feature Engineering

```

data_outlier['total_guest'] = data_outlier['adults'] + data_outlier['children'] + data_outlier['babies']

```

total nights	0.0s
--------------	------

```

data_outlier['total_nights'] = data_outlier['stays_in_weekend_nights'] + data_outlier['stays_in_week_nights']

```

revenue	0.0s
---------	------

```

data_outlier['revenue'] = data_outlier['adr'] * data_outlier['total_nights']

```

revenue	0.0s
---------	------

Untuk meningkatkan kemampuan prediktif model, dilakukan feature engineering yang menciptakan tiga fitur gabungan baru dengan relevansi bisnis tinggi. Pertama, 'total\_guest' yang mengintegrasikan jumlah semua pengunjung (adults + children + babies) memberikan gambaran lengkap tentang ukuran kelompok tamu, yang berpotensi mempengaruhi tingkat pembatalan karena koordinasi rombongan yang lebih besar cenderung lebih rumit. Kedua, 'total\_nights' yang menggabungkan 'stays\_in\_weekend\_nights' dan 'stays\_in\_week\_nights' mewakili durasi total menginap, memfasilitasi analisis korelasi antara panjangnya menginap dengan kemungkinan pembatalan. Ketiga, fitur 'revenue' yang dihitung dari perkalian 'adr' (average daily rate) dengan 'total\_nights' menghasilkan estimasi nilai finansial total dari setiap pemesanan, memungkinkan model untuk mempelajari apakah pemesanan

dengan nilai yang lebih tinggi memiliki pola pembatalan yang berbeda. Fitur-fitur derivatif ini memperkaya representasi data dan memungkinkan model menangkap interaksi kompleks antar variabel yang mungkin terlewatkan oleh fitur individual.

#### g. Encoding Categorical Data

```
Label Encoder

le_cancel = LabelEncoder()
le_notcancel = LabelEncoder()

data_outlier['prev_cancel_cat_le'] = le_cancel.fit_transform(data_outlier['prev_cancel_cat'])
data_outlier['prev_notcancel_cat_le'] = le_notcancel.fit_transform(data_outlier['prev_notcancel_cat'])

# Setelah ini, kamu bisa drop kolom aslinya jika hanya ingin pakai versi LE
data_outlier = data_outlier.drop(columns=['prev_cancel_cat', 'prev_notcancel_cat'])

✓ 0.0s
```

```
Value Mapping

month_map = {
    'January': 1, 'February': 2, 'March': 3, 'April': 4,
    'May': 5, 'June': 6, 'July': 7, 'August': 8,
    'September': 9, 'October': 10, 'November': 11, 'December': 12
}
data_outlier['arrival_date_month_num'] = data_outlier['arrival_date_month'].map(month_map)
data_outlier = data_outlier.drop(columns=['arrival_date_month'])

✓ 0.0s
```

Dalam penanganan variabel kategorikal, implementasi dilakukan dengan pendekatan hybrid yang memanfaatkan kekuatan algoritma CatBoost dan pertimbangan struktur semantik data. Dari 13 kolom kategorikal yang teridentifikasi, hanya beberapa yang memerlukan encoding manual karena karakteristik khususnya. Untuk fitur 'arrival\_date\_month', dilakukan value mapping untuk mengkonversi nama bulan menjadi representasi numerik berurutan (1-12), mempertahankan informasi siklik dan urutan kronologis waktu. Sementara itu, fitur 'prev\_cancel\_cat' dan 'prev\_notcancel\_cat' diubah dengan LabelEncoder karena memiliki karakteristik ordinal yang menunjukkan tingkatan frekuensi pembatalan atau non-pembatalan sebelumnya. Untuk 10 fitur kategorikal lainnya (termasuk 'hotel', 'meal', 'country', dll.), tidak dilakukan encoding manual karena akan ditangani langsung oleh CatBoost melalui parameter 'cat\_features', yang merupakan keunggulan utama algoritma ini dalam menangani data kategorikal tanpa one-hot encoding yang dapat menyebabkan dimensionalitas tinggi.

#### h. Train & Test Split

```

# Pisahkan fitur dan target
X = data_outlier.drop(columns=['is_canceled'])
y = data_outlier['is_canceled']

# Train-test split (misal 80:20)
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)

```

✓ 0.0s

Dalam tahap pembagian dataset untuk pelatihan dan pengujian model, stratified train-test split dengan rasio 80:20 diimplementasikan untuk memastikan representasi yang seimbang dari kedua kelas target. Fitur input (X) didefinisikan dengan mengeluarkan kolom target 'is\_canceled' dari dataset, sementara vektor target (y) berisi nilai-nilai kolom tersebut. Parameter 'stratify=y' memastikan proporsi kelas yang sama antara set pelatihan dan pengujian, terlihat dari distribusi label yang identik pada kedua set: 62.46% untuk Label 0 (pemesanan yang tidak dibatalkan) dan 37.54% untuk Label 1 (pemesanan yang dibatalkan). Pendekatan ini sangat penting untuk menghindari bias sampling, terutama dalam kasus imbalanced class seperti ini, karena memastikan model dilatih dan dievaluasi pada distribusi kelas yang konsisten. Penggunaan 'random\_state=42' memberikan hasil split yang deterministik, memungkinkan reproducibility dalam pengembangan dan evaluasi model.

### i. Data Standardization (StandardScaler)

```

num_cols = X_train.select_dtypes(include=['int64', 'float64']).columns.tolist()
scaler = StandardScaler()
X_train[num_cols] = scaler.fit_transform(X_train[num_cols])
X_test[num_cols] = scaler.transform(X_test[num_cols])

# (opsional) tampilkan 5 baris pertama hasil scaling
print("\nContoh hasil standardisasi fitur numerik:")
X_train[num_cols].head()

```

✓ 0.1s

Standardisasi data dilakukan khusus untuk fitur-fitur numerik dalam dataset menggunakan teknik StandardScaler dari scikit-learn, yang mentransformasi variabel ke distribusi dengan mean=0 dan standard deviation=1. Proses ini dimulai dengan identifikasi kolom bertipe 'int64' dan 'float64' dari X\_train, dilanjutkan dengan fitting scaler pada data latih dan transformasi pada kedua set data (train dan test). Metode ini sangat penting untuk memastikan semua fitur numerik berada pada skala yang sebanding, mencegah variabel dengan rentang nilai besar mendominasi algoritma pembelajaran. Penerapan standardisasi hanya pada fitur numerik (bukan kategorikal) merupakan praktik tepat karena fitur kategorikal akan ditangani secara terpisah oleh CatBoost. Penting dicatat bahwa scaler hanya di-fit pada data latih untuk mencegah data leakage, kemudian parameter scaler yang sama diterapkan untuk

mentransformasi data uji, memastikan konsistensi transformasi antara kedua set data dan integritas evaluasi model.

### 6.1.2. Modelling (CatBoost Classifier)

#### a. Build & Train Model

```

cat_features = [
    'hotel', 'meal', 'country', 'market_segment', 'distribution_channel',
    'reserved_room_type', 'assigned_room_type', 'deposit_type', 'customer_type', 'lead_time_bin']
    ✓ 0.0s

# Buat model CatBoostClassifier dengan parameter default
catboost_model = cb.CatBoostClassifier()

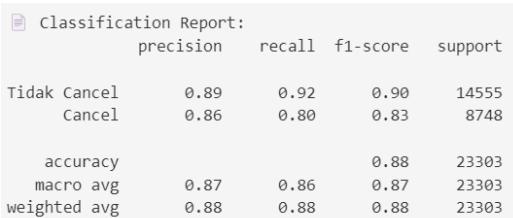
# Training model, hanya perlu cat_features untuk fitur kategorikal
catboost_model.fit(
    X_train, y_train,
    cat_features=cat_features,
    eval_set=(X_test, y_test),
    verbose=50,          # Output setiap 50 iterasi
    use_best_model=True  # Pilih model terbaik di validation
)

# Prediksi pada data test
y_pred = catboost_model.predict(X_test)
y_pred_proba = catboost_model.predict_proba(X_test)[:, 1]
    ✓ 1m 58.4s

```

Model prediksi pembatalan hotel diimplementasikan menggunakan algoritma CatBoost, pilihan optimal untuk dataset yang memiliki kombinasi fitur numerik dan kategorikal. Proses dimulai dengan mendefinisikan 10 fitur kategorikal ('hotel', 'meal', 'country', dll.) yang akan ditangani secara native oleh CatBoost tanpa encoding tambahan. Model dilatih dengan parameter default dan learning rate 0.097133, dengan validasi berkelanjutan menggunakan data test sebagai eval\_set. Log pelatihan menunjukkan penurunan konsisten pada loss function dari 0.61 pada iterasi awal hingga 0.27 pada iterasi ke-999, dengan model terbaik dicapai pada iterasi ke-994. Fitur 'use\_best\_model=True' memastikan model final merupakan versi dengan performa terbaik pada data validasi. Proses ini membutuhkan waktu sekitar 2 menit (1m 58.4s) untuk 1000 iterasi, menunjukkan efisiensi komputasi yang baik. Setelah pelatihan, model digunakan untuk menghasilkan baik prediksi kelas (y\_pred) maupun probabilitas pembatalan (y\_pred\_proba) pada data test, memberikan fleksibilitas dalam interpretasi dan evaluasi.

#### b. Model Evaluation

Metric Evaluation	Interpretasi																														
 <table border="1"> <thead> <tr> <th></th> <th>precision</th> <th>recall</th> <th>f1-score</th> <th>support</th> </tr> </thead> <tbody> <tr> <td>Tidak Cancel</td> <td>0.89</td> <td>0.92</td> <td>0.90</td> <td>14555</td> </tr> <tr> <td>Cancel</td> <td>0.86</td> <td>0.80</td> <td>0.83</td> <td>8748</td> </tr> <tr> <td>accuracy</td> <td></td> <td></td> <td>0.88</td> <td>23303</td> </tr> <tr> <td>macro avg</td> <td>0.87</td> <td>0.86</td> <td>0.87</td> <td>23303</td> </tr> <tr> <td>weighted avg</td> <td>0.88</td> <td>0.88</td> <td>0.88</td> <td>23303</td> </tr> </tbody> </table>		precision	recall	f1-score	support	Tidak Cancel	0.89	0.92	0.90	14555	Cancel	0.86	0.80	0.83	8748	accuracy			0.88	23303	macro avg	0.87	0.86	0.87	23303	weighted avg	0.88	0.88	0.88	23303	<p>Model CatBoost menunjukkan performa klasifikasi yang sangat baik dengan akurasi 0.88 pada data pengujian. Secara detail, model ini unggul dalam prediksi kelas "Tidak Cancel" (precision 0.89, recall 0.92, F1-score 0.90) dan juga menunjukkan performa yang solid untuk kelas "Cancel" (precision 0.86, recall 0.80, F1-score 0.83). Baik macro</p>
	precision	recall	f1-score	support																											
Tidak Cancel	0.89	0.92	0.90	14555																											
Cancel	0.86	0.80	0.83	8748																											
accuracy			0.88	23303																											
macro avg	0.87	0.86	0.87	23303																											
weighted avg	0.88	0.88	0.88	23303																											

	<p>average maupun weighted average menunjukkan konsistensi dengan F1-score 0.87-0.88, yang mengindikasikan model memiliki keseimbangan baik antara precision dan recall pada kedua kelas meskipun adanya slight imbalance dalam dataset (14,555 data "Tidak Cancel" vs 8,748 data "Cancel"). Hasil ini menunjukkan model memiliki kapabilitas prediktif yang kuat untuk mengidentifikasi pembatalan reservasi hotel.</p>													
<table border="1"> <thead> <tr> <th colspan="2" rowspan="2"></th> <th colspan="2">Predicted Label</th> </tr> <tr> <th>Tidak Cancel</th> <th>Cancel</th> </tr> </thead> <tbody> <tr> <th rowspan="2">Actual Label</th> <th>Tidak Cancel</th> <td>13445</td> <td>1110</td> </tr> <tr> <th>Cancel</th> <td>1736</td> <td>7012</td> </tr> </tbody> </table>			Predicted Label		Tidak Cancel	Cancel	Actual Label	Tidak Cancel	13445	1110	Cancel	1736	7012	<p>Confusion matrix menunjukkan model CatBoost berhasil dengan benar memprediksi 13,445 kasus tidak pembatalan (True Negative) dan 7,012 kasus pembatalan (True Positive). Terdapat 1,736 kasus false negative (pembatalan yang diprediksi sebagai tidak pembatalan) dan 1,110 kasus false positive (tidak pembatalan yang diprediksi sebagai pembatalan). Distribusi ini mengkonfirmasi performa yang baik dalam kedua kelas, dengan slightly better performance untuk kelas "Tidak Cancel".</p>
			Predicted Label											
		Tidak Cancel	Cancel											
Actual Label	Tidak Cancel	13445	1110											
	Cancel	1736	7012											
<p>The ROC curve plot shows the performance of the CatBoost model. The x-axis is the False Positive Rate (FPR) ranging from 0.0 to 1.0. The y-axis is the True Positive Rate (TPR) ranging from 0.0 to 1.0. A solid blue curve represents the ROC Curve for the model, starting at (0,0) and ending at (1,1), indicating an AUC of 0.95. A dashed diagonal line represents the Random Guess line, where TPR equals FPR.</p>	<p>Kurva ROC menunjukkan performa model CatBoost yang sangat baik dengan nilai AUC (Area Under the Curve) mencapai 0.95. Nilai yang mendekati 1.0 ini mengindikasikan kemampuan diskriminasi model yang sangat tinggi dalam membedakan antara kelas pembatalan dan tidak pembatalan. Bentuk kurva yang curam di bagian awal dan tinggi di atas garis diagonal (Random Guess) menunjukkan model memiliki sensitivitas tinggi dan spesifisitas yang baik pada berbagai threshold keputusan, menegaskan reliabilitas model dalam memprediksi pembatalan reservasi hotel.</p>													

### 6.1.3. Hyperparameter Tuning dengan Optuna

- a. Implementasi Bayesian Optimization

```

# Fungsi objective untuk Optuna
def objective(trial):
    params = {
        "iterations": trial.suggest_int("iterations", 200, 1000),
        "learning_rate": trial.suggest_float("learning_rate", 0.01, 0.3, log=True),
        "depth": trial.suggest_int("depth", 4, 10),
        "l2_leaf_reg": trial.suggest_float("l2_leaf_reg", 1, 10),
        "random_strength": trial.suggest_float("random_strength", 1e-9, 10, log=True),
        "bagging_temperature": trial.suggest_float("bagging_temperature", 0.0, 1.0),
        "border_count": trial.suggest_int("border_count", 32, 255),
        "verbose": False,
        "cat_features": cat_features,
        "task_type": "GPU",
        "early_stopping_rounds": 30,
    }
    model = cb.CatBoostClassifier(**params)
    model.fit(
        X_train, y_train,
        eval_set=(X_test, y_test),
        use_best_model=True,
    )
    y_pred_proba = model.predict_proba(X_test)[:, 1]
    score = roc_auc_score(y_test, y_pred_proba)
    ...return score

# Optuna study
study = optuna.create_study(direction="maximize", study_name="CatBoost Optuna")
study.optimize(objective, n_trials=50, show_progress_bar=True)

print("Best trial:")
print(f" AUC: {study.best_value:.4f}")
print(" Params:", study.best_params)

```

✓ 45m 8.3s

Best trial:

- AUC: 0.9571
- Params: {'iterations': 869, 'learning\_rate': 0.06929599922255182, 'depth': 10, 'l2\_leaf\_reg': 2.868324534345466, 'random\_strength': 0.014775484327958601, 'bagging\_temperature': 0.37466937706827463, 'border\_count': 245}

Proses hyperparameter tuning menggunakan framework Optuna dilakukan untuk mengoptimalkan model CatBoost, dengan metric AUC sebagai objective yang dimaksimalkan. Dalam fungsi objective, Optuna melakukan pencarian parameter optimal pada ruang yang telah didefinisikan: iterations (200-1000), learning\_rate (0.01-0.3 dalam skala logaritmik), depth (4-10), l2\_leaf\_reg (1-10), random\_strength (1e-9-10 dalam skala logaritmik), bagging\_temperature (0-1), dan border\_count (32-255). Setelah 50 trials yang memakan waktu 45 menit, trial terbaik (ke-35) menghasilkan peningkatan performa signifikan dengan AUC 0.9571 menggunakan konfigurasi: 869 iterasi, learning rate 0.069, kedalaman pohon 10, regularisasi L2 2.87, random strength 0.015, bagging temperature 0.37, dan border count 245. Konfigurasi optimal ini menghasilkan model dengan kemampuan diskriminasi yang lebih baik dibandingkan model baseline.

## b. Implementasi Model dengan Parameter Optimal

```

# Dapatkan parameter terbaik dari hasil Optuna
best_params = study.best_params.copy()
best_params["cat_features"] = cat_features
best_params["task_type"] = "GPU" # atau "CPU" jika tidak ada GPU
best_params["verbose"] = 50
best_params["early_stopping_rounds"] = 30

# Buat dan fit ulang model CatBoost dengan parameter terbaik
best_model = cb.CatBoostClassifier(**best_params)
best_model.fit(
    X_train, y_train,
    eval_set=(X_test, y_test),
    use_best_model=True
)

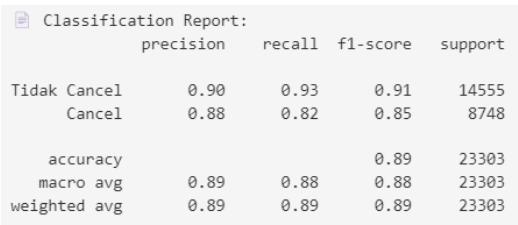
y_pred = best_model.predict(X_test)
y_pred_proba = best_model.predict_proba(X_test)[:, 1]

```

✓ 1m 46.2s

Setelah mendapatkan parameter terbaik dari proses tuning Optuna, model CatBoost diimplementasikan ulang dengan konfigurasi optimal tersebut. Log pelatihan menunjukkan penurunan loss yang lebih cepat dan stabil: dari 0.61 pada iterasi awal hingga 0.25 pada iterasi ke-868, dengan model terbaik tercapai pada iterasi ke-862. Implementasi ini menetapkan `cat_features` untuk penanganan fitur kategorikal, memanfaatkan akselerasi GPU untuk pelatihan yang lebih cepat, serta mengaktifkan early stopping dengan parameter 30 rounds untuk mencegah overfitting. Proses pelatihan membutuhkan waktu 1 menit 46 detik, menghasilkan model final yang dioptimalkan untuk performa prediktif maksimal. Setelah pelatihan, model digunakan untuk menghasilkan prediksi kelas dan probabilitas pada data test, yang kemudian akan digunakan untuk evaluasi performa final.

### c. Model Evaluation Setelah Tuning

Metric Evaluation	Interpretasi
 <pre> Classification Report: precision    recall   f1-score   support Tidak Cancel      0.90      0.93      0.91     14555     Cancel        0.88      0.82      0.85      8748  accuracy           -         -         -       23303 macro avg        0.89      0.88      0.88     23303 weighted avg      0.89      0.89      0.89     23303 </pre>	<p>Model CatBoost menunjukkan performa klasifikasi yang sangat baik dengan akurasi 0.88 pada data pengujian. Secara detail, model ini unggul dalam prediksi kelas "Tidak Cancel" (precision 0.89, recall 0.92, F1-score 0.90) dan juga menunjukkan performa yang solid untuk kelas "Cancel" (precision 0.86, recall 0.80, F1-score 0.83). Baik macro average maupun weighted average menunjukkan konsistensi dengan F1-score 0.87-0.88, yang mengindikasikan model memiliki keseimbangan baik antara precision dan recall pada kedua kelas meskipun adanya slight imbalance dalam dataset (14,555 data "Tidak Cancel" vs 8,748 data "Cancel"). Hasil ini menunjukkan</p>

	<p>model memiliki kapabilitas prediktif yang kuat untuk mengidentifikasi pembatalan reservasi hotel.</p>														
<p>Confusion Matrix - CatBoost</p> <table border="1"> <thead> <tr> <th colspan="2" rowspan="2">Actual Label</th> <th colspan="2">Predicted Label</th> </tr> <tr> <th>Tidak Cancel</th> <th>Cancel</th> </tr> </thead> <tbody> <tr> <th rowspan="2">Tidak Cancel</th> <td>13564</td> <td>991</td> </tr> <tr> <td>1565</td> <td>7183</td> </tr> <tr> <th>Cancel</th> <td></td> <td></td> </tr> </tbody> </table>	Actual Label		Predicted Label		Tidak Cancel	Cancel	Tidak Cancel	13564	991	1565	7183	Cancel			<p>Perbandingan classification report menunjukkan peningkatan performa yang konsisten setelah hyperparameter tuning. Akurasi model meningkat dari 0.88 menjadi 0.89. Untuk kelas "Tidak Cancel", precision meningkat dari 0.89 menjadi 0.90 dan recall naik dari 0.92 menjadi 0.93, menghasilkan peningkatan F1-score dari 0.90 menjadi 0.91. Pada kelas "Cancel", peningkatan lebih signifikan dengan precision naik dari 0.86 menjadi 0.88 dan recall naik dari 0.80 menjadi 0.82, meningkatkan F1-score dari 0.83 menjadi 0.85. Baik macro average maupun weighted average juga meningkat dari 0.87-0.88 menjadi 0.88-0.89, menunjukkan model hasil tuning memiliki kemampuan prediksi yang lebih baik dan lebih seimbang pada kedua kelas.</p>
Actual Label			Predicted Label												
		Tidak Cancel	Cancel												
Tidak Cancel	13564	991													
	1565	7183													
Cancel															
<p>ROC Curve - CatBoost</p> <p>True Positive Rate</p> <p>False Positive Rate</p> <p>ROC Curve (AUC = 0.96)</p> <p>Random Guess</p>	<p>Kurva ROC model setelah hyperparameter tuning menunjukkan peningkatan performa dengan nilai AUC meningkat dari 0.95 menjadi 0.96. Peningkatan ini mengindikasikan model hasil tuning memiliki kemampuan diskriminasi yang lebih tinggi dalam membedakan antara kelas pembatalan dan tidak pembatalan. Kurva tampak lebih menjauhi garis diagonal dan lebih curam di bagian awal, menunjukkan peningkatan sensitivitas dengan meminimalkan false positive rate. Hasil ini membuktikan bahwa proses optimasi parameter dengan Optuna berhasil meningkatkan kemampuan model dalam mengklasifikasikan pembatalan reservasi hotel, membuat model lebih handal untuk implementasi di sistem prediksi pembatalan real.</p>														

## 6.2. Klusterisasi (K-Means)

Klusterisasi (clustering) adalah salah satu teknik dalam *unsupervised learning* yang digunakan untuk mengelompokkan data ke dalam beberapa kelompok (kluster) berdasarkan kemiripan atau kedekatan karakteristik tertentu. Berbeda dengan *classification* yang memerlukan label data, klusterisasi tidak membutuhkan label sebelumnya, melainkan mencari pola alami dalam data.

Algoritma K-Means bekerja dengan menentukan jumlah kluster (K) di awal, lalu mengelompokkan data berdasarkan jarak terdekat ke titik pusat kluster (*centroid*) yang terus diperbarui hingga proses konvergen. Model ini sangat cocok untuk dataset pemesanan hotel yang memiliki kombinasi fitur numerik dan kategorikal (yang telah diproses), karena dapat mengungkapkan segmentasi pelanggan secara alami, seperti membedakan antara tamu korporat, keluarga, atau wisatawan individu. Penggunaan K-Means dalam penelitian ini bertujuan untuk membantu manajemen hotel dalam memahami pola reservasi pelanggan dan perilaku pembatalan secara kelompok.

### 6.2.1 Pre-Modelling

#### a. Pemilihan Fitur untuk Clustering

```
# Pilih beberapa fitur numerik & kategorikal
cols_clustering = [
    "lead_time",
    "stays_in_weekend_nights",
    "stays_in_week_nights",
    "adults",
    "children",
    "babies",
    "adr",
    "customer_type",      # kategorikal
    "deposit_type"        # kategorikal
]

# Ambil data yang sudah diproses outliernya
data_for_cluster = data_outlier[cols_clustering].copy()

# One-hot encoding untuk kolom kategorikal
categorical_cols = ["customer_type", "deposit_type"]
data_clustering = pd.get_dummies(data_for_cluster, columns=categorical_cols, drop_first=True)
```

Pemilihan fitur untuk proses klusterisasi dilakukan dengan menggabungkan beberapa variabel numerik dan kategorikal yang dianggap relevan. fitur numerik yang digunakan meliputi lead\_time, stays\_in\_weekend\_nights, stays\_in\_week\_nights, adults, children, babies, dan adr, sedangkan fitur kategorikal mencakup customer\_type dan deposit\_type. Dataset yang digunakan merupakan data yang telah melalui proses pembersihan outlier sebelumnya, sehingga memastikan bahwa input untuk klusterisasi bebas dari nilai ekstrim yang dapat mengganggu pembentukan pola.

Proses one-hot encoding terhadap fitur kategorikal menggunakan fungsi pd.get\_dummies() dari pustaka pandas. Tujuannya adalah mengubah data kategorikal menjadi bentuk numerik biner agar kompatibel dengan algoritma klusterisasi seperti K-Means yang hanya menerima input numerik. Parameter drop\_first=True digunakan untuk menghindari dummy trap dan menjaga interpretabilitas model. Hasil akhir dari tahapan ini adalah data yang telah terstandardisasi secara struktural dan siap digunakan dalam proses klusterisasi untuk menemukan segmentasi pelanggan hotel berdasarkan perilaku pemesanannya.

b. Data Standardization ( StandardScaler )

```
# Daftar kolom numerik (sebelum encoding)
numerical_cols = [
    "lead_time",
    "stays_in_weekend_nights",
    "stays_in_week_nights",
    "adults",
    "children",
    "babies",
    "adr"
]

# Pisahkan fitur numerik dan dummy kategorikal
data_num = data_clustering[numerical_cols]
data_dummy = data_clustering.drop(columns=numerical_cols)

# Scaling hanya fitur numerik
scaler = StandardScaler()
data_num_scaled = pd.DataFrame(scaler.fit_transform(data_num), columns=numerical_cols, index=data_clustering.index)

# Gabungkan Lagi
data_clustering_scaled = pd.concat([data_num_scaled, data_dummy], axis=1)
```

Proses standardisasi dilakukan dengan menggunakan StandardScaler dari pustaka sklearn, yang mengubah data numerik agar memiliki nilai rata-rata 0 dan standar deviasi 1. Transformasi ini menghasilkan fitur numerik yang distandarisasi dan lebih seimbang dalam kontribusinya saat penghitungan jarak antar data. Setelah proses scaling, data numerik yang telah ditransformasikan dikonversi kembali ke dalam bentuk DataFrame dengan mempertahankan indeks dan nama kolom aslinya. Penggabungan dilakukan secara horizontal (axis=1) untuk menghasilkan dataset akhir yang siap digunakan dalam algoritma klusterisasi. Pendekatan ini memastikan bahwa baik fitur numerik maupun kategorikal dapat berkontribusi secara optimal dalam pembentukan kluster yang representatif terhadap pola perilaku pelanggan hotel.

c. Data Reduction Using PCA with 2 Components

```
pca = PCA(n_components=2)
df_pca = pca.fit_transform(data_clustering_scaled)

pca_df = pd.DataFrame(df_pca, columns=['PCA1', 'PCA2'])
pca_df.head()
```

PCA diterapkan untuk menurunkan dimensi data dari sejumlah besar fitur menjadi dua komponen utama, yaitu PCA1 dan PCA2. Proses dimulai dengan parameter n\_components=2, yang berarti akan diambil dua komponen utama dengan varian terbesar. fungsi fit\_transform() kemudian digunakan untuk menghitung dan menerapkan transformasi PCA terhadap data yang telah diskalakan (data\_clustering\_scaled). Hasil dari transformasi ini adalah matriks baru yang berisi nilai representatif dari dua dimensi utama seperti gambar dibawah berikut.

	PCA1	PCA2
0	-1.430846	-1.389314
1	-0.620821	-2.680574
2	-1.898225	0.134916
3	-1.885921	0.115302
4	-0.842132	0.539955

#### d. Determining the Optimal K (Elbow Method)

```
# Menentukan rentang nilai k (jumlah cluster yang dicoba)
inertia = []
k_range = range(1, 11)

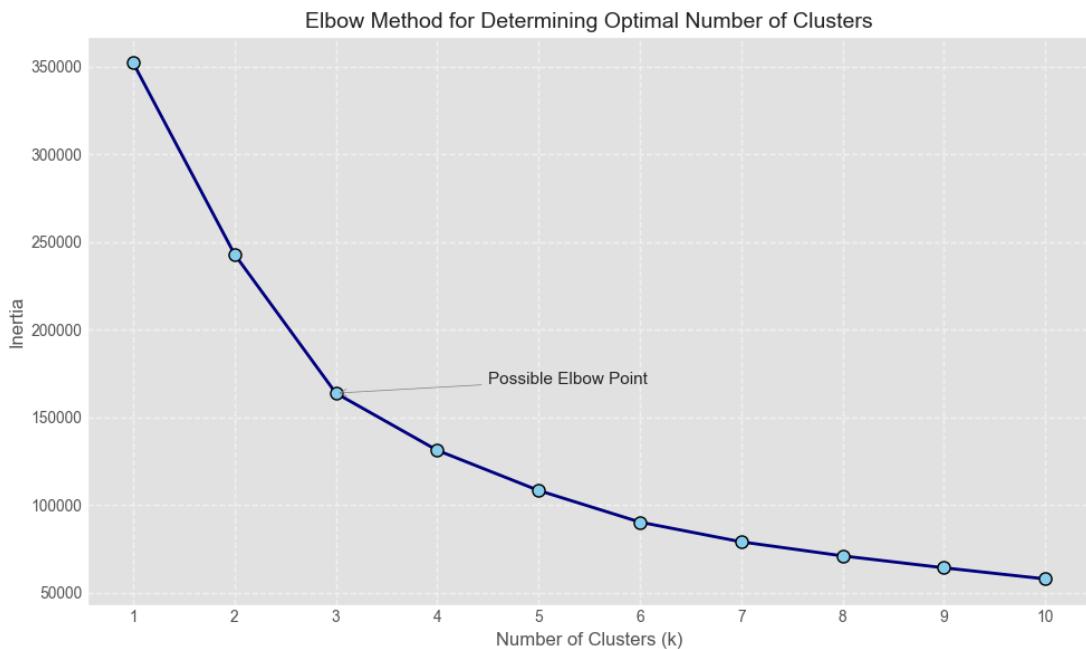
# Melakukan KMeans untuk setiap nilai k dan menyimpan inertia-nya
for k in k_range:
    kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
    kmeans.fit(pca_df) # Menggunakan data PCA yang sudah di-scale
    inertia.append(kmeans.inertia_)

# Visualisasi menggunakan Elbow Method
plt.figure(figsize=(10, 6))
plt.plot(k_range, inertia, marker='o', linestyle='-', color='navy', linewidth=2, markersize=8, markerfacecolor='skyblue', markeredgecolor='black')
plt.xticks(k_range)
plt.xlabel('Number of Clusters (k)', fontsize=12)
plt.ylabel('Inertia', fontsize=12)
plt.title('Elbow Method for Determining Optimal Number of Clusters', fontsize=14)
plt.grid(True, linestyle='--', alpha=0.5)

# Menambahkan anotasi untuk meng-highlight titik elbow (jika terlihat)
plt.annotate('Possible Elbow Point',
            xy=(3, inertia[2]),
            xytext=(4.5, inertia[2] + 5000),
            arrowprops=dict(arrowstyle='->', color='gray'),
            fontsize=11)

plt.tight_layout()
plt.show()
```

Untuk menentukan jumlah kluster yang optimal dalam algoritma K-Means, digunakan Elbow Method, yaitu teknik visual untuk mengidentifikasi titik optimal (elbow point) di mana penambahan jumlah kluster tidak lagi menghasilkan penurunan yang signifikan terhadap nilai inertia. Inertia sendiri mengukur total jarak kuadrat antar data ke centroid klusternya masing-masing—semakin kecil nilai inertia, semakin baik pemisahan kluster. Untuk setiap nilai K, model K-Means dibangun dan dilatih pada data hasil PCA (pca\_df), kemudian nilai inertia disimpan dalam list inertia. Parameter random\_state=42 dan n\_init=10 digunakan untuk menjaga kestabilan hasil. Setelah seluruh nilai inertia diperoleh, dilakukan visualisasi menggunakan matplotlib untuk menampilkan grafik hubungan antara jumlah kluster (K) dan nilai inertia.



Grafik tersebut digunakan untuk mengidentifikasi titik elbow, yaitu titik yang menunjukkan perubahan penurunan inertia yang tidak lagi signifikan. Dalam kode, ditambahkan anotasi untuk menandai titik elbow secara manual dengan label “Possible Elbow Point”. Titik ini diasumsikan sebagai jumlah kluster optimal berdasarkan lekukan grafik. Hasil visualisasi ini menjadi dasar penting dalam memilih jumlah kluster yang akan digunakan pada tahap klusterisasi berikutnya.

### 6.3.2 Build K-Means Model

```
# Nilai k optimal berdasarkan analisis sebelumnya
optimal_k = 3

# Melakukan clustering dengan K-Means
kmeans = KMeans(n_clusters=optimal_k, random_state=42, n_init=10)
cluster_labels = kmeans.fit_predict(pca_df)

# Menambahkan label cluster ke dataframe original
data_outlier['Cluster'] = cluster_labels

# Menghitung silhouette score keseluruhan
silhouette_avg = silhouette_score(pca_df, cluster_labels)
print(f"Silhouette Score untuk k={optimal_k}: {silhouette_avg:.4f}")
```

Silhouette Score adalah metrik yang mengukur seberapa baik setiap titik data cocok dengan klusternya sendiri dibandingkan dengan kluster lain. Nilai ini berada dalam rentang -1 hingga 1, di mana nilai mendekati 1 menunjukkan pemisahan kluster yang baik. Pada analisis ini, **nilai silhouette score** yang diperoleh untuk k=3 adalah **0.4343**, yang menunjukkan kualitas klusterisasi yang cukup baik dan mengindikasikan bahwa

pemilihan jumlah kluster sudah cukup tepat untuk memisahkan data berdasarkan pola yang terbentuk.

### a. Visualizing Clustering Results and Centroids

```
# Melakukan clustering KMeans pada data PCA
optimal_k = 3
kmeans = KMeans(n_clusters=optimal_k, random_state=42, n_init=10)
cluster_labels = kmeans.fit_predict(pca_df)

# Tambahkan Label cluster ke data outlier
data_outlier['Cluster'] = cluster_labels

# Hitung silhouette score
silhouette_avg = silhouette_score(pca_df, cluster_labels)

# Mendapatkan centroid di ruang asli dan transform ke PCA
centroids = kmeans.cluster_centers_
centroids_pca = centroids # Cluster sudah di-fit pada PCA, jadi centroid sudah di ruang PCA

# Visualisasi
plt.figure(figsize=(10, 8))
palette = sns.color_palette("viridis", as_cmap=True)

# Plot data per cluster
scatter = plt.scatter(
    pca_df.iloc[:, 0], pca_df.iloc[:, 1],
    c=cluster_labels, cmap='viridis',
    s=60, alpha=0.75, edgecolor='k', linewidth=0.5,
    label='Data'
)

# Plot centroid
plt.scatter(
    centroids_pca[:, 0], centroids_pca[:, 1],
    c='red', s=320, marker='X', edgecolor='black', linewidth=2,
    label='Centroids'
)

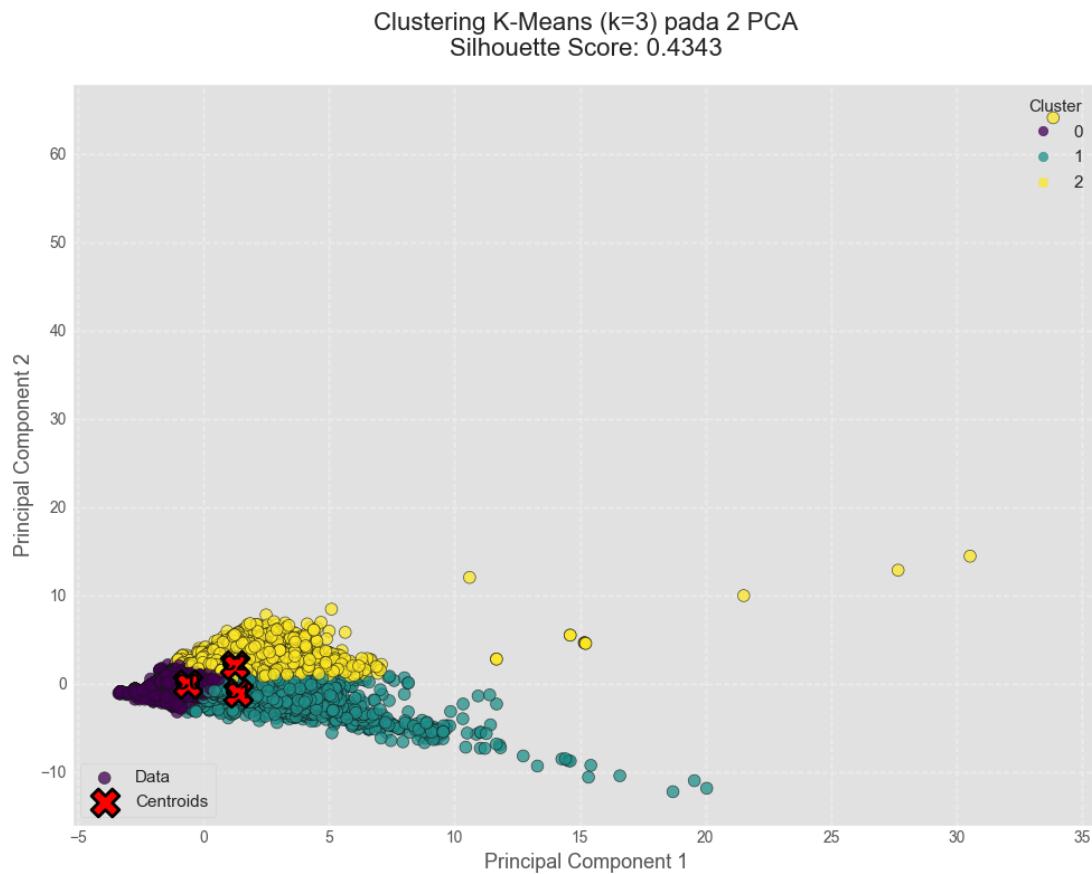
# Anotasi centroid (opsional, bisa dihapus)
for i, (x, y) in enumerate(centroids_pca):
    plt.text(x, y, f'C{i}', color='black', fontsize=14, fontweight='bold', ha='center', va='center')

# Title & label
plt.title(f'Clustering K-Means (k={optimal_k}) pada 2 PCA\nSilhouette Score: {silhouette_avg:.4f}', fontsize=16, pad=18)
plt.xlabel('Principal Component 1', fontsize=13)
plt.ylabel('Principal Component 2', fontsize=13)

# Legend
legend1 = plt.legend(*scatter.legend_elements(), title="Cluster", loc="upper right", fontsize=11)
plt.gca().add_artist(legend1)
plt.legend(["Data", "Centroids"], loc='lower left', fontsize=11, frameon=True)

plt.grid(True, linestyle='--', alpha=0.35)
plt.tight_layout()
plt.show()
```

Setiap titik pada grafik merepresentasikan satu data pemesanan hotel yang telah ditransformasikan ke dalam dua komponen utama PCA, yaitu Principal Component 1 dan Principal Component 2. Model K-Means kemudian memetakan setiap data ke dalam salah satu dari tiga kluster, yang ditunjukkan dengan warna berbeda. Selain menampilkan distribusi data, visualisasi ini juga menunjukkan posisi centroid dari masing-masing kluster, yang ditandai dengan simbol “X” besar berwarna merah. Centroid ini merupakan titik pusat dari masing-masing kluster yang dihitung oleh K-Means. Koordinat centroid telah ditransformasikan ke ruang PCA agar dapat divisualisasikan bersama dengan data.



Hasil visualisasi tersebut mencantumkan **Silhouette Score** sebesar **0.4343**, yang menggambarkan kualitas pemisahan antar kluster. Nilai ini mendekati 0.5, menunjukkan bahwa kluster yang terbentuk sudah cukup terpisah namun masih ada area tumpang tindih. Secara keseluruhan, visualisasi ini memberikan gambaran yang jelas tentang hasil klusterisasi, membantu dalam menganalisis pola kelompok pelanggan hotel berdasarkan data historis pemesanan, dan dapat digunakan sebagai dasar segmentasi dalam pengambilan keputusan strategis.

### 6.3.3 Export K-Means Model

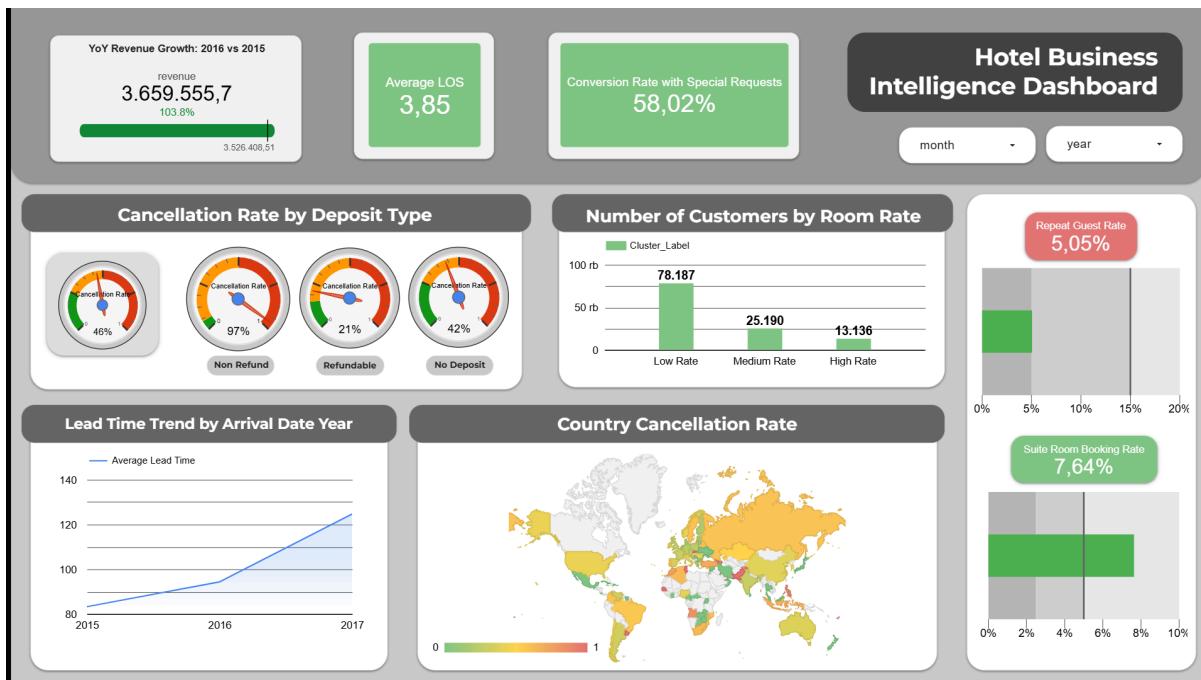
```
import joblib
joblib.dump(kmeans, "kmeans.pkl")
```

Joblib merupakan salah satu library di Python yang umum digunakan untuk menyimpan dan memuat objek model machine learning. Perintah `joblib.dump(kmeans, "kmeans.pkl")` berfungsi untuk menyimpan objek model K-Means ke dalam file dengan format .pkl (pickle). File tersebut nantinya dapat dimuat kembali untuk keperluan prediksi atau analisis lanjutan tanpa perlu melatih ulang model dari awal.

## BAB VII PERANCANGAN DASHBOARD KPI

### 7.1. Implementasi Dashboard Visualisasi KPI

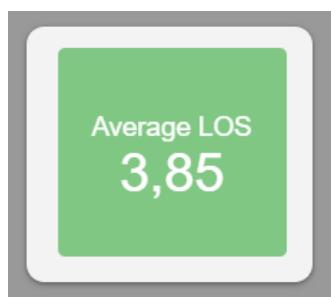
Setelah merancang indikator kinerja utama (KPI) dari keempat perspektif balance scorecard, seluruh KPI tersebut diimplementasikan dalam bentuk visualisasi interaktif menggunakan Looker Studio. Pemilihan jenis chart disesuaikan dengan karakteristik metrik dan kebutuhan analisis, agar informasi dapat disampaikan secara efektif dan mudah dipahami oleh pengguna.



[Link Dashboard](#)

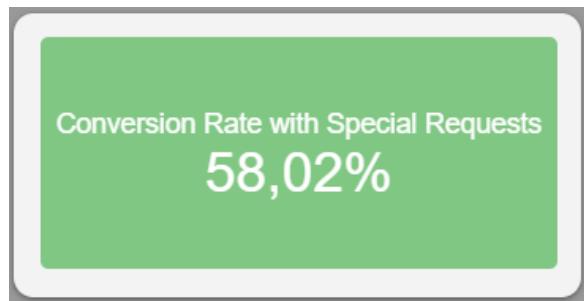
Berikut adalah penjelasan implementasi visualisasi dari masing - masing KPI:

- KPI 1: Average Length of Stay (LOS) - Text



Visualisasi ini menampilkan persentase pada rata-rata jumlah malam menginap per reservasi pada customer yang tidak dibatalkan. Nilai ditampilkan dalam bentuk teks agar dapat terbaca langsung secara eksplisit. Metrik ini membantu hotel dalam perhitungan untuk setiap operasionalnya.

- KPI 2: Special Request Fulfillment Rate - Text



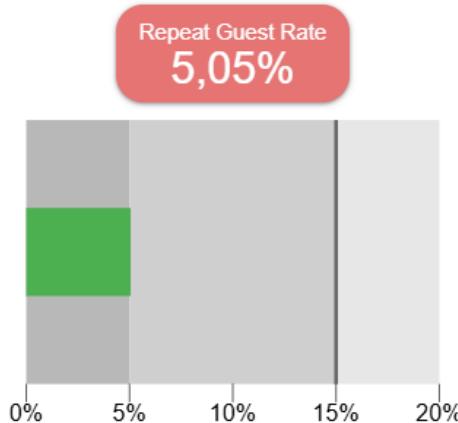
Visualisasi ini menampilkan persentase pada jumlah permintaan khusus yang dipenuhi per reservasi dan persentase reservasi dengan permintaan khusus yang tidak dibatalkan. Nilai ditampilkan dalam bentuk teks agar dapat terbaca langsung secara eksplisit. Metrik ini membantu hotel dalam perhitungan untuk setiap operasionalnya terutama pada permintaan khusus pada tamu hotel.

- KPI 3: Year-over-Year Revenue Growth - Text



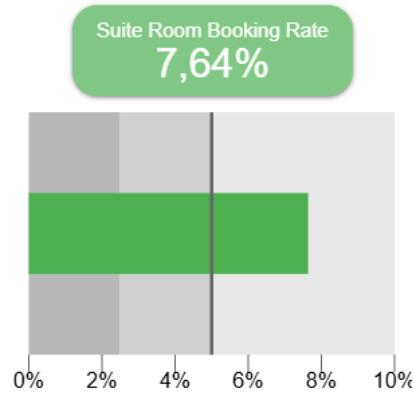
Visualisasi ini menampilkan persentase pada mengukur pertumbuhan bisnis hotel secara keseluruhan dan efektivitas strategi revenue management yang diterapkan. Nilai ditampilkan dalam bentuk teks agar dapat terbaca langsung secara eksplisit. Metrik ini membantu hotel dalam perhitungan keuntungan yang didapat pada setiap operasionalnya.

- KPI 4: Repeat Guest Rate - Bullet Chart



Visualisasi dalam bentuk bullet chart ini menampilkan persentase tamu yang kembali menginap di hotel dalam kurun waktu tertentu dalam bentuk persentase. Visual ini menunjukkan nilai aktual, target, dan benchmark dalam satu baris yang efisien.

- KPI 5: Suite Booking Rate - Bullet Chart



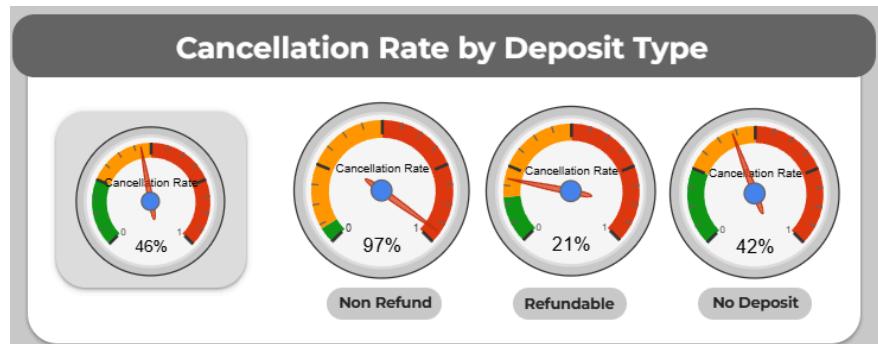
Visualisasi dalam bentuk bullet chart ini menampilkan persentase pemesanan untuk kamar tipe Suite (kode "G") dari total seluruh pemesanan hotel. Visual ini menunjukkan nilai aktual, target, dan benchmark dalam satu baris yang efisien. Selain itu juga, metrik ini untuk mengukur keberhasilan hotel dalam menjual kategori kamar premium yang memiliki margin profit tertinggi.

- KPI 6: Cancellation Rate - Gauge Chart



Visualisasi dalam bentuk gauge chart ini menampilkan persentase reservasi yang dibatalkan dari total reservasi yang dilakukan. Cara kerja visual ini dengan membandingkan hasil dengan batas maksimal (100%). Gauge chart digunakan bertujuan untuk memantau dan mengendalikan tingkat pembatalan, manajemen dapat mengambil langkah preventif untuk meminimalkan kerugian.

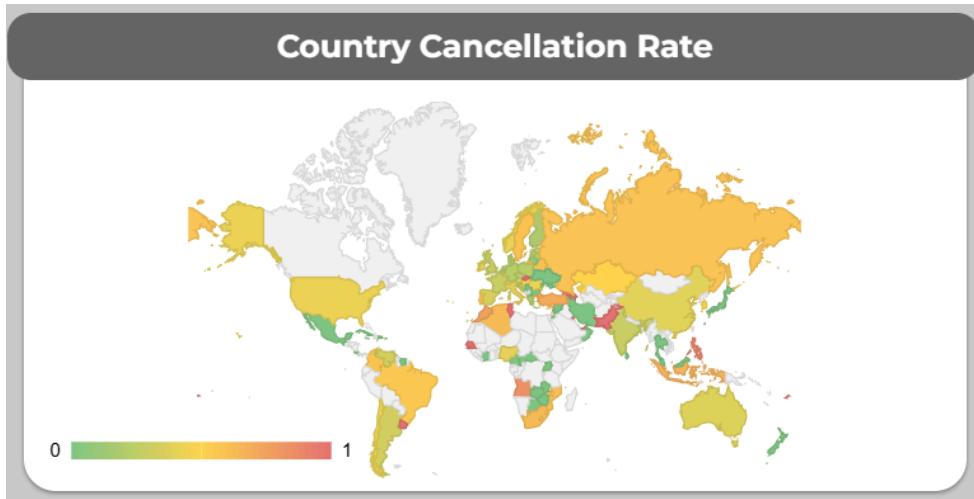
- KPI 7: Cancellation Rate by Deposit Type - Gauge Chart



Visualisasi dalam bentuk gauge chart ini menampilkan persentase efektivitas kebijakan deposit dalam mencegah pembatalan, ditunjukkan dengan persentase pembatalan untuk setiap tipe deposit. Cara kerja visual ini dengan membandingkan hasil (Non Refund, Refundable, No

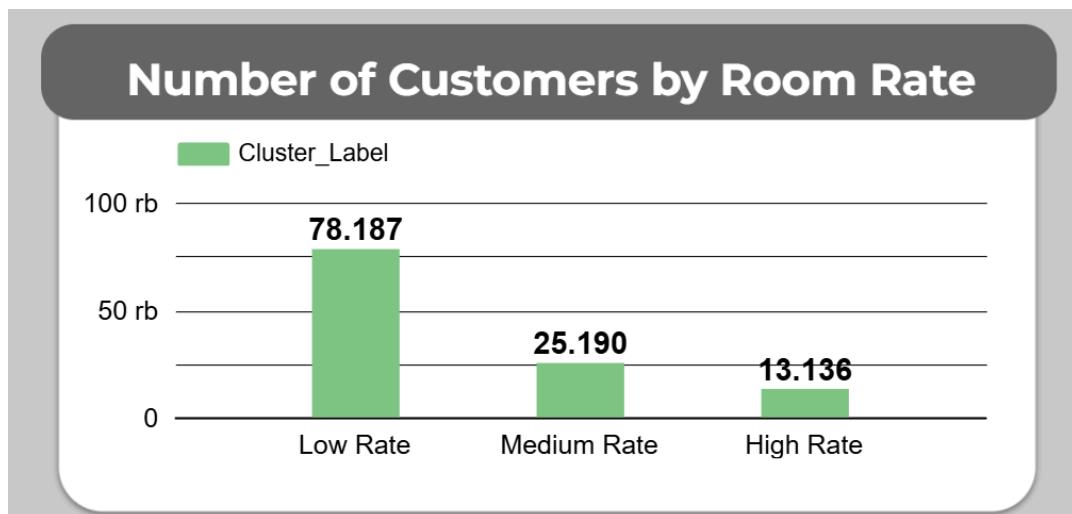
Deposit) dengan batas maksimal (100%). Gauge chart digunakan agar pengguna dapat melihat apakah rata - rata pembatalan masih dalam batas aman atau justru terlalu besar.

- KPI 8: Country Cancellation Rate - Map Chart



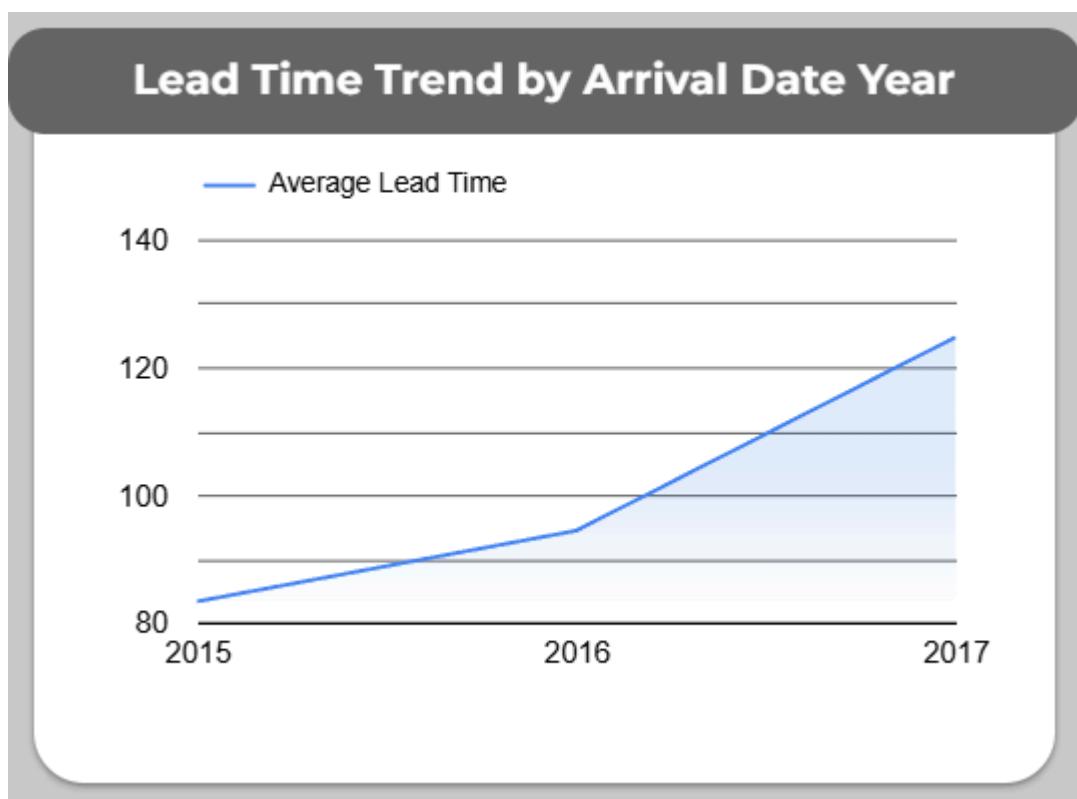
Visualisasi peta (Map Chart) digunakan sebagai pemantauan persentase reservasi yang dibatalkan dari total reservasi untuk setiap negara asal tamu. Visualisasi geografis memungkinkan pembaca langsung melihat wilayah dengan pembatalan pemesanan hotel dengan tingkat tertinggi maupun terendah.

- KPI 9 : Number of Customers by Room Rate - Bar Chart



Visualisasi bar chart ini digunakan untuk menampilkan jumlah pelanggan pada setiap kategori tarif kamar (Low Rate, Medium Rate, dan High Rate). Visualisasi ini membantu mengidentifikasi preferensi pelanggan terhadap tarif kamar, sehingga manajemen dapat menyesuaikan strategi harga, promosi, dan segmentasi pasar untuk mengoptimalkan okupansi serta pendapatan hotel.

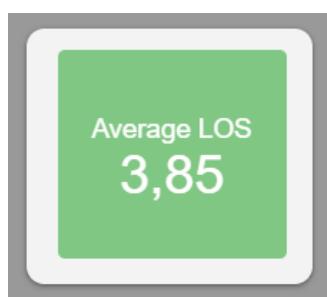
- KPI 10: Lead Time Trend by Arrival Date Year - Line Chart



Visualisasi line chart digunakan untuk menunjukkan tren rata-rata lead time (hari sebelum kedatangan) berdasarkan tahun kedatangan tamu (arrival\_date\_year). Grafik ini memberikan gambaran fluktuasi kedatangan tamu dari waktu ke waktu dan digunakan untuk menganalisis pengaruh dalam strategi kebijakan pembatalan dan deposit.

## 7.2. Analisis & Evaluasi KPI

### 7.2.1 Average Length of Stay (LOS)



#### Evaluasi Terhadap Target

Aspek	Detail
Nilai Aktual	3,85 malam
Target	$\geq 2$ malam

Status	Hijau (Melebihi Target)
Selisih dengan Target	+1,85 malam

### Kekuatan

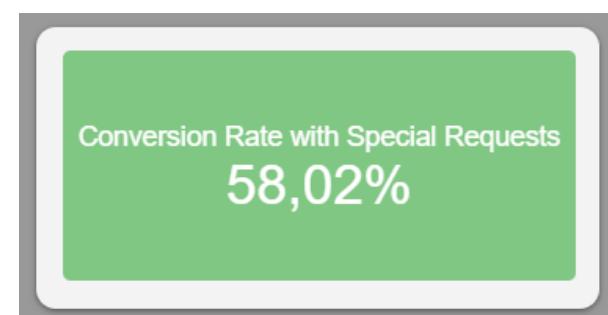
1. Kinerja Sangat Baik: LOS aktual 3,85 malam hampir dua kali lipat dari target minimal 2 malam, menunjukkan keberhasilan strategi yang signifikan dalam mendorong tamu menginap lebih lama.
2. Implikasi Finansial Positif: Dengan rata-rata menginap yang jauh melampaui target, hotel mendapatkan beberapa keuntungan:
  - Pengurangan biaya operasional per malam karena penurunan frekuensi check-in/check-out
  - Peningkatan pendapatan total per tamu
  - Kesempatan upselling yang lebih besar untuk layanan tambahan
3. Efisiensi Operasional: Housekeeping dan front desk memiliki beban kerja yang lebih efisien dengan pergantian kamar yang lebih rendah per total malam terjual.

### Rekomendasi Strategis

Meskipun KPI ini sudah melampaui target dengan sangat baik, berikut beberapa rekomendasi untuk mempertahankan dan meningkatkan performa:

1. Analisis Segmentasi: Lakukan analisis mendalam untuk mengidentifikasi segmen tamu mana yang berkontribusi paling tinggi terhadap LOS, dan fokuskan upaya pemasaran pada segmen tersebut.
2. Pengembangan Paket: Ciptakan lebih banyak paket menginap dengan durasi yang lebih panjang (4-7 hari) dengan insentif tambahan.
3. Pertimbangkan Revisi Target: Dengan performa aktual yang jauh melampaui target, pertimbangkan untuk menaikkan target menjadi 3 atau 3,5 malam untuk tahun mendatang.
4. Program Loyalitas yang Ditingkatkan: Kembangkan benefit khusus untuk tamu yang menginap lebih lama dalam program loyalitas hotel.
5. Analisis Musiman: Pantau LOS secara musiman untuk memastikan konsistensi sepanjang tahun, atau mengidentifikasi periode yang memerlukan strategi khusus.

#### 7.2.2 Special Request Fulfillment Rate



## Evaluasi Terhadap Target

Aspek	Detail
Nilai Aktual	58,02%
Target	$\geq 50\%$
Status	Hijau (Melebihi Target)
Selisih dengan Target	+8,02%

## Kekuatan

1. Performa di Atas Target: Special Request Fulfillment Rate telah melampaui target minimal 50% dengan nilai aktual 58,02%, menunjukkan komitmen hotel yang kuat terhadap pemenuhan kebutuhan khusus tamu.
2. Indikator Kepuasan Pelanggan: Tingginya tingkat pemenuhan permintaan khusus berkontribusi positif terhadap pengalaman tamu secara keseluruhan, yang berpotensi meningkatkan skor kepuasan dan ulasan positif.
3. Differensiasi Layanan: Kemampuan memenuhi lebih dari setengah permintaan khusus menjadi nilai tambah yang membedakan hotel dari kompetitor dan mendorong loyalitas pelanggan.

## Rekomendasi Strategis

1. Identifikasi Pola Permintaan: Lakukan analisis mendalam terhadap jenis permintaan khusus yang paling sering diminta dan yang paling sering dipenuhi untuk mengoptimalkan alokasi sumber daya.
2. Pelatihan Staf Terarah: Berikan pelatihan khusus pada staf terkait permintaan spesial yang paling umum, memastikan mereka memiliki kemampuan dan sumber daya untuk memenuhinya dengan efisien.
3. Komunikasi Proaktif: Implementasikan sistem notifikasi pra-kedatangan untuk mengkonfirmasi permintaan khusus dan menyampaikan status pemenuhannya sebelum tamu melakukan check-in.
4. Peningkatan Target: Dengan performa saat ini, pertimbangkan untuk meningkatkan target menjadi 65% untuk tahun berikutnya, mendorong peningkatan layanan yang berkelanjutan.

### 7.2.3 Year-over-Year Revenue Growth



#### Evaluasi Terhadap Target

Aspek	Detail
Nilai Aktual	103,8%
Target	≥ 3%
Status	Hijau (Melebihi Target)
Selisih dengan Target	+0,8%

#### Kekuatan

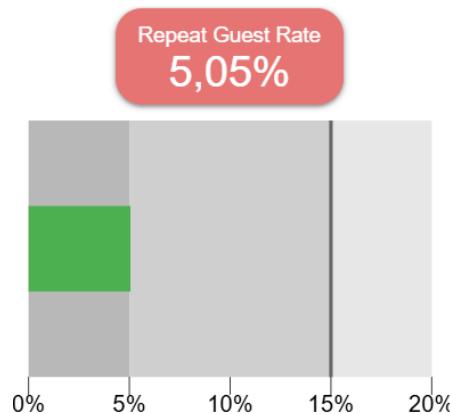
1. Pertumbuhan di Atas Target: Year-over-Year Revenue Growth mencapai 3,8%, melampaui target minimal 3% yang telah ditetapkan, menunjukkan performa keuangan yang positif.
2. Peningkatan Pendapatan Absolut: Terjadi kenaikan pendapatan sebesar 133.147,19 dari tahun 2015 ke 2016, mengindikasikan peningkatan aktivitas bisnis yang signifikan.
3. Posisi Kompetitif yang Kuat: Dengan pertumbuhan di atas target yang juga melebihi rata-rata industri, hotel berhasil memperkuat posisinya di pasar perhotelan.

#### Rekomendasi Strategis

1. Analisis Sumber Pertumbuhan: Lakukan kajian mendalam terhadap sumber-sumber utama pertumbuhan pendapatan, apakah berasal dari peningkatan tingkat hunian, kenaikan harga kamar, atau pendapatan non-kamar.
2. Strategi Pricing Dinamis: Implementasikan strategi harga yang lebih dinamis berdasarkan analisis permintaan musiman dan segmen pasar untuk meningkatkan pendapatan per kamar yang tersedia (RevPAR).
3. Diversifikasi Pendapatan: Kembangkan lebih banyak sumber pendapatan non-kamar seperti F&B, layanan spa, atau event untuk meningkatkan total pendapatan dan mengurangi ketergantungan pada pendapatan kamar.

- Target yang Lebih Ambisius: Dengan pencapaian yang melebihi target saat ini, pertimbangkan untuk menetapkan target pertumbuhan yang lebih tinggi untuk tahun berikutnya, misalnya 4-5%.

#### 7.2.4 Repeat Guest Rate



#### Evaluasi Terhadap Target

Aspek	Detail
Nilai Aktual	5,05%
Target	$\geq 15\%$
Status	Merah (Di bawah Target)
Selisih dengan Target	-9,95%

#### Kekuatan

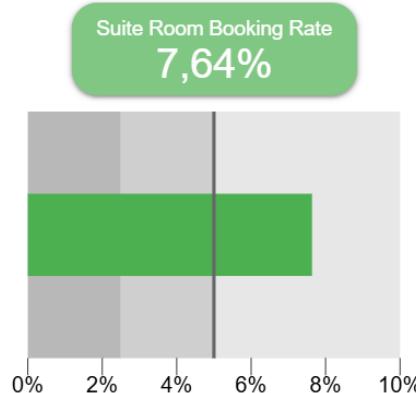
- Basis Pengukuran yang Jelas: KPI ini memiliki metode perhitungan yang jelas dan dapat ditelusuri, memungkinkan pengukuran yang konsisten dan akurat dari waktu ke waktu.
- Potensi Peningkatan Signifikan: Nilai aktual yang masih rendah menunjukkan adanya ruang besar untuk peningkatan dan potensi pertumbuhan pendapatan yang belum dimanfaatkan.
- Kesadaran akan Pentingnya Loyalitas: Adanya KPI ini menunjukkan bahwa hotel telah menyadari pentingnya loyalitas pelanggan dalam strategi bisnis jangka panjang.

#### Rekomendasi Strategis

- Program Loyalitas yang Komprehensif: Kembangkan atau tingkatkan program loyalitas yang menawarkan nilai konkret kepada tamu yang kembali, seperti akumulasi poin, upgrade gratis, atau layanan khusus.

2. Personalisasi Pengalaman Tamu: Manfaatkan data dari kunjungan sebelumnya untuk mempersonalisasi pengalaman tamu yang kembali, menunjukkan bahwa preferensi mereka diingat dan dihargai.
3. Komunikasi Pasca menginap yang Terarah: Implementasikan strategi email marketing yang terarah untuk tetap berhubungan dengan tamu setelah check-out, termasuk penawaran khusus untuk kunjungan berikutnya.
4. Analisis Segmentasi Tamu: Lakukan analisis untuk mengidentifikasi segmen tamu dengan potensi tertinggi untuk menjadi pelanggan berulang dan fokuskan upaya retensi pada segmen-segmen tersebut.
5. Pengembangan Staf: Latih staf untuk mengenali dan memberikan perhatian khusus kepada tamu yang kembali, menciptakan pengalaman yang membuat mereka merasa dihargai.
6. Kajian Benchmarking: Lakukan perbandingan dengan standar industri untuk memahami apakah target 15% realistik untuk segmen pasar dan lokasi hotel spesifik Anda.

### 7.2.5 Suite Booking Rate



#### Evaluasi Terhadap Target

Aspek	Detail
Nilai Aktual	7,64%
Target	$\geq 5\%$
Status	Hijau (Melebihi Target)
Selisih dengan Target	+2,64%

#### Kekuatan

1. Performa Jauh di Atas Target: Suite Booking Rate mencapai 7,64%, lebih dari 50% di atas target minimal 5%, menunjukkan keberhasilan strategi penjualan kamar premium.
2. Dampak Positif pada ADR: Tingginya tingkat pemesanan suite berkontribusi signifikan terhadap peningkatan Average Daily Rate (ADR) hotel, meningkatkan profitabilitas secara keseluruhan.
3. Strategi Upselling yang Efektif: Pencapaian ini mengindikasikan bahwa strategi upselling dan cross-selling yang diterapkan telah berhasil mendorong tamu untuk memilih akomodasi kelas yang lebih tinggi.

### **Rekomendasi Strategis**

1. Optimalisasi Harga Suite: Evaluasi strategi penetapan harga untuk kamar suite, dengan mempertimbangkan peluang untuk sedikit meningkatkan harga tanpa mengurangi tingkat pemesanan yang sudah tinggi.
2. Segmentasi Pemasaran: Analisis karakteristik tamu yang cenderung memesan kamar suite dan kembangkan kampanye pemasaran yang lebih terarah untuk segmen tersebut.
3. Pengalaman Suite yang Ditingkatkan: Investasikan dalam peningkatan fasilitas dan layanan di kamar suite untuk meningkatkan nilai yang dirasakan tamu, mendorong ulasan positif, dan meningkatkan pemesanan berulang.
4. Paket Suite yang Menarik: Kembangkan paket menginap khusus untuk kamar suite yang menyertakan layanan nilai tambah seperti sarapan gratis, spa, atau kredit F&B untuk meningkatkan daya tarik dan nilai keseluruhan.
5. Target yang Lebih Ambisius: Dengan pencapaian jauh di atas target, pertimbangkan untuk menaikkan target Suite Booking Rate menjadi 8% untuk tahun berikutnya untuk mendorong peningkatan performa yang berkelanjutan.

#### **7.2.6 Cancellation Rate**



#### **Evaluasi Terhadap Target**

<b>Aspek</b>	<b>Detail</b>
Nilai Aktual	46%

Target	$\leq 15\%$
Status	Kuning (Di Atas Target, Berada dalam Zona Kuning)
Selisih dengan Target	+31%

### Kekuatan

1. Tingkat Pembatalan Sangat Tinggi: Cancellation Rate sebesar 46% jauh melebihi target maksimal 15%, mengindikasikan masalah serius dalam retensi pemesanan.
2. Dampak pada Perencanaan Operasional: Tingginya angka pembatalan menyulitkan prediksi okupansi yang akurat, berpotensi mengganggu perencanaan staf, persediaan, dan sumber daya lainnya.
3. Risiko Pendapatan: Hampir setengah dari semua pemesanan dibatalkan, menciptakan ketidakpastian signifikan pada proyeksi pendapatan dan memaksa hotel untuk lebih bergantung pada pemesanan last-minute.

### Rekomendasi Strategis

1. Evaluasi Kebijakan Pembatalan: Pertimbangkan untuk menerapkan kebijakan pembatalan yang lebih ketat, seperti biaya pembatalan bertingkat yang meningkat mendekati tanggal check-in atau deposit non-refundable yang lebih besar.
2. Analisis Pola Pembatalan: Identifikasi faktor-faktor utama yang berkontribusi terhadap tingginya tingkat pembatalan, seperti segmen pasar tertentu, saluran pemesanan, durasi menginap, atau periode musiman.
3. Strategi Overbooking yang Terkalibrasi: Implementasikan strategi overbooking yang diperhitungkan dengan cermat berdasarkan data historis pembatalan untuk mengoptimalkan tingkat hunian aktual.
4. Program Insentif Komitmen: Kembangkan program yang memberikan insentif kepada tamu untuk berkomitmen pada reservasi mereka, seperti diskon untuk pembayaran di muka atau layanan tambahan untuk pemesanan non-refundable.
5. Komunikasi Proaktif: Tingkatkan komunikasi dengan tamu sebelum kedatangan untuk mengkonfirmasi reservasi dan mengatasi potensi masalah yang mungkin menyebabkan pembatalan.

### 7.2.7 Cancellation Rate by Deposit Type



#### Evaluasi Terhadap Target

Tipe Deposit	Nilai Aktual	Target	Status	Selisih dengan Target
Non-Refundable	97%	$\leq 5\%$	Merah (Jauh di atas Target)	+92%
Refundable	21%	$\leq 15\%$	Kuning (Di atas Target)	+6%
No Deposit	42%	$\leq 25\%$	Kuning (Di atas Target)	+17%

#### Kekuatan

##### Non-Refundable (97%)

1. Performa Sangat Mengkhawatirkan: Tingkat pembatalan 97% pada pemesanan Non-Refundable sangat mengkhawatirkan dan kontra-intuitif, mengingat tipe deposit ini seharusnya memiliki tingkat pembatalan terendah.
2. Kemungkinan Masalah Sistemik: Tingkat pembatalan yang sangat tinggi ini mengindikasikan kemungkinan adanya masalah dalam:
  - Kebijakan pembatalan yang tidak diterapkan dengan benar
  - Pengembalian dana yang mudah diperoleh meski berstatus non-refundable
  - Kesalahan dalam kategorisasi atau pencatatan data

##### Refundable (21%)

1. Performa Kurang Optimal: Tingkat pembatalan 21% untuk pemesanan Refundable berada di atas target maksimal 15%, mengindikasikan area yang memerlukan perbaikan.
2. Potensi Perbaikan: Selisih 6% dari target menunjukkan adanya peluang perbaikan yang relatif lebih mudah dicapai dibandingkan tipe deposit lainnya.

##### No Deposit (42%)

1. Tingkat Pembatalan Tinggi: Tingkat pembatalan 42% untuk pemesanan tanpa deposit cukup tinggi, meskipun memang diharapkan lebih tinggi dibandingkan tipe deposit lainnya.
2. Selisih Signifikan dengan Target: Berada 17% di atas target maksimal 25% mengindikasikan perlunya evaluasi pendekatan terhadap pemesanan tanpa deposit.

## **Rekomendasi Strategis**

### Non-Refundable

1. Audit Sistem: Lakukan audit menyeluruh terhadap kebijakan dan implementasi pemesanan Non-Refundable untuk memastikan tidak ada kesalahan dalam penerapan atau pencatatan data.
2. Evaluasi Kebijakan: Tinjau kebijakan pembatalan Non-Refundable untuk memastikan konsistensi dalam penerapan dan komunikasi kepada tamu.
3. Persyaratan yang Lebih Ketat: Pertimbangkan untuk menerapkan persyaratan pembayaran penuh di muka untuk memperkuat komitmen pemesanan.

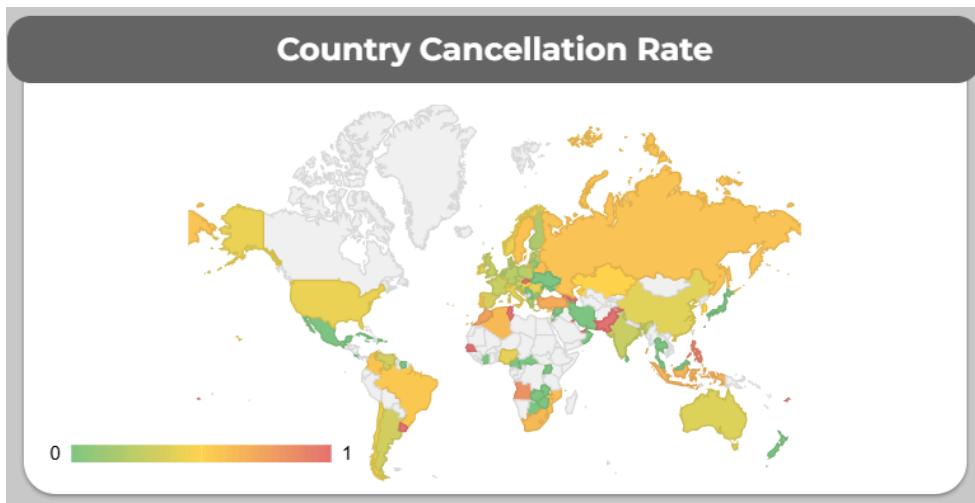
### Refundable

1. Insentif untuk Mempertahankan Pemesanan: Tawarkan insentif seperti upgrade kamar atau layanan tambahan gratis jika tamu tidak membatalkan pemesanan Refundable.
2. Kebijakan Pembatalan Bertingkat: Implementasikan kebijakan pembatalan bertingkat di mana biaya pembatalan meningkat seiring mendekati tanggal check-in.

### No Deposit

1. Sistem Konfirmasi Proaktif: Implementasikan sistem konfirmasi otomatis berkala untuk pemesanan tanpa deposit untuk mengurangi pembatalan mendadak.
2. Deposit Nominal: Pertimbangkan memperkenalkan deposit nominal (bukan nol) untuk menciptakan komitmen finansial minimal tanpa menghilangkan daya tarik opsi "No Deposit".
3. Perbaikan Komunikasi: Tingkatkan komunikasi pra-kedatangan untuk pemesanan tanpa deposit dengan mengirimkan informasi yang menarik tentang hotel dan tujuan untuk mempertahankan antusiasme tamu.

### 7.2.8 Country Cancellation Rate



#### Kekuatan

1. Didukung oleh hasil model prediksi: Variabel country memiliki nilai feature importance tertinggi (~13,5), menunjukkan bahwa negara asal tamu sangat berpengaruh terhadap kemungkinan pembatalan reservasi.
2. Membantu strategi pemasaran yang lebih tepat sasaran: Dengan mengetahui negara mana yang memiliki tingkat pembatalan tinggi, hotel dapat menyusun kebijakan pemasaran, promosi, atau penyesuaian layanan berdasarkan karakteristik tamu dari tiap negara.
3. Mendukung kebijakan reservasi yang lebih efektif: Data ini dapat dijadikan dasar untuk menentukan kebijakan deposit atau fleksibilitas pemesanan sesuai dengan risiko pembatalan di masing-masing negara.
4. Visualisasi mudah dipahami: Penyajian dalam bentuk *map chart* memudahkan identifikasi pola pembatalan berdasarkan wilayah geografis, sehingga mempermudah analisis dan pengambilan keputusan secara cepat.
5. Mencapai Target (Hijau): Selandia Baru, beberapa negara Eropa Barat, dan beberapa negara Asia Tenggara kemungkinan memiliki tingkat pembatalan di bawah target 25%.
6. Mendekati Target (Kuning): Amerika Utara, Australia, dan beberapa negara Eropa lainnya kemungkinan memiliki tingkat pembatalan mendekati target 25%.
7. Di Atas Target (Oranye-Merah): Beberapa negara di Timur Tengah, Afrika, dan Asia kemungkinan memiliki tingkat pembatalan jauh di atas target 25%.

#### Rekomendasi Strategis

Untuk Negara dengan Performa Baik (Hijau)

- Analisis Praktik Terbaik: Identifikasi faktor yang berkontribusi terhadap tingkat pembatalan rendah dari negara-negara ini (misalnya pola pemesanan, preferensi tamu, saluran pemesanan).

- Replikasi Strategi: Terapkan strategi pemasaran dan kebijakan reservasi serupa yang telah berhasil di negara-negara ini ke negara lain dengan karakteristik pasar yang mirip.
- Penguatan Hubungan: Pertahankan dan perkuat hubungan dengan agen perjalanan dan mitra distribusi di negara-negara ini untuk mempertahankan performa yang baik.

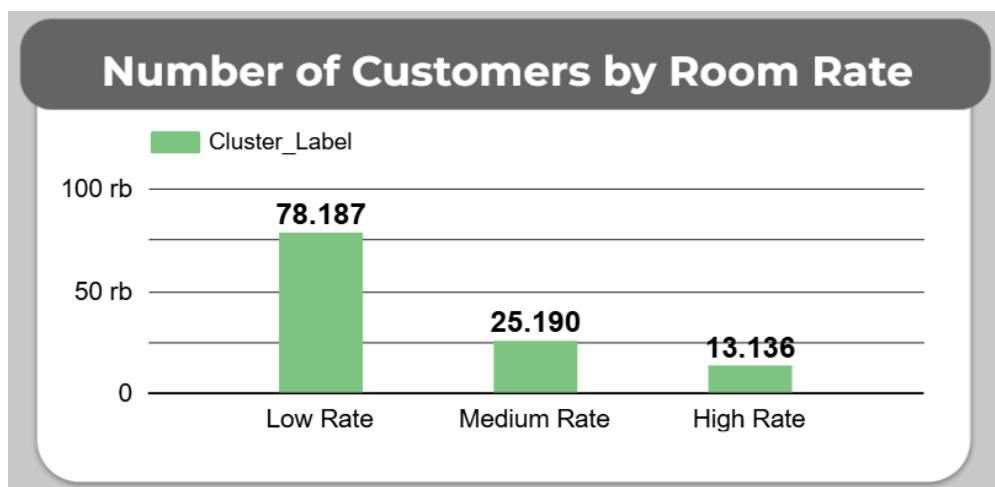
Untuk Negara dengan Performa Moderat (Kuning)

- Incentif Terpersonalisasi: Kembangkan incentif khusus untuk tamu dari negara-negara ini yang mendorong mereka untuk mempertahankan pemesanan, seperti upgrade atau layanan tambahan.
- Komunikasi Ditingkatkan: Implementasikan komunikasi pra-kedatangan yang ditingkatkan dan disesuaikan dengan preferensi budaya untuk tamu dari negara-negara ini.
- Kebijakan Pembatalan yang Dioptimalkan: Sesuaikan kebijakan pembatalan untuk menyeimbangkan fleksibilitas dan komitmen berdasarkan pola pemesanan di negara-negara ini.

Untuk Negara dengan Performa Buruk (Oranye-Merah)

- Kebijakan Deposit yang Direvisi: Pertimbangkan kebijakan deposit yang lebih ketat untuk pemesanan dari negara-negara dengan tingkat pembatalan tinggi, seperti deposit non-refundable yang lebih besar.
- Program Edukasi: Kembangkan materi informatif dan program edukasi untuk tamu dari negara-negara ini tentang kebijakan pembatalan dan implikasinya.
- Analisis Akar Masalah: Lakukan penelitian mendalam untuk mengidentifikasi penyebab tingginya tingkat pembatalan dari negara-negara ini (misalnya masalah visa, keterjangkauan, atau faktor budaya).
- Strategi Saluran Pemesanan: Evaluasi dan potensial restrukturisasi saluran pemesanan yang digunakan untuk negara-negara dengan tingkat pembatalan tinggi.

#### 7.2.9 Number of Customers by Room Rate



## Evaluasi Terhadap Target

Aspek	Detail
Nilai Aktual	67% pelanggan di kategori Low Rate
Target	$\geq 50\%$ pelanggan di kategori Low Rate
Selisih dengan Target	+17%

## Analisis Distribusi Tarif

1. Dominasi Kategori Low Rate (67%)
  - Mayoritas pelanggan (67%) memilih kamar dengan tarif rendah, menunjukkan sensitivitas harga yang tinggi di segmen pasar utama hotel
  - Pencapaian ini jauh melebihi target minimal 50%, mengindikasikan strategi penetapan harga yang berhasil menarik volume pelanggan tinggi
2. Segmen Medium Rate (21.6%)
  - Sekitar seperlima pelanggan memilih kategori harga menengah
  - Mewakili segmen pelanggan yang mencari keseimbangan antara harga dan nilai
3. Segmen High Rate (11.3%)
  - Segmen premium menyumbang porsi terkecil namun tetap signifikan
  - Meskipun jumlahnya lebih sedikit, segmen ini kemungkinan berkontribusi signifikan terhadap total pendapatan karena margin yang lebih tinggi

## Kekuatan

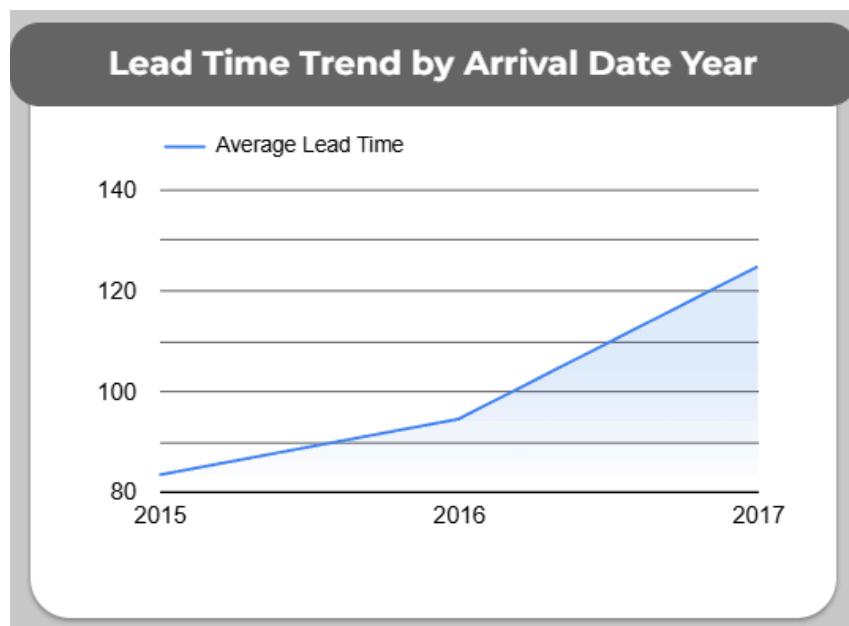
1. Strategi Volume yang Efektif: Keberhasilan menarik mayoritas pelanggan ke kategori Low Rate menunjukkan strategi penetapan harga yang efektif untuk memaksimalkan okupansi.
2. Diversifikasi Segmen: Distribusi pelanggan di tiga kategori tarif menunjukkan kemampuan hotel untuk melayani berbagai segmen pasar dengan kebutuhan dan anggaran yang berbeda.
3. Basis Kuat untuk Upselling: Volume tinggi di kategori Low Rate memberikan peluang besar untuk strategi upselling dan cross-selling untuk meningkatkan pendapatan per pelanggan.
4. Daya Tarik Pasar yang Tinggi: Pencapaian 67% pelanggan di kategori Low Rate (melebihi target 50%) menunjukkan daya tarik pasar yang kuat dan posisi kompetitif yang baik.

## Rekomendasi Strategis

1. Optimalisasi Revenue Mix
  - a. Meskipun telah mencapai target volume untuk Low Rate, analisis lebih lanjut diperlukan untuk memastikan bahwa distribusi ini optimal untuk total pendapatan

- b. Evaluasi kontribusi pendapatan dari masing-masing segmen untuk memastikan profitabilitas keseluruhan
- 2. Program Upselling Terarah
  - a. Kembangkan strategi upselling yang efektif untuk mengkonversi sebagian pelanggan Low Rate menjadi Medium Rate
  - b. Fokus pada komunikasi nilai tambah untuk mendorong peningkatan kategori
- 3. Pengembangan Segmen Premium
  - a. Identifikasi peluang untuk meningkatkan proporsi pelanggan High Rate tanpa mengorbankan volume keseluruhan
  - b. Tingkatkan proposisi nilai untuk segmen premium untuk menarik lebih banyak pelanggan yang kurang sensitif terhadap harga
- 4. Strategi Dinamis Berdasarkan Musim
  - a. Implementasikan strategi penetapan harga yang lebih dinamis yang menyesuaikan proporsi target untuk kategori tarif berdasarkan musim dan tingkat permintaan
  - b. Tingkatkan fleksibilitas untuk mengoptimalkan pendapatan selama periode permintaan tinggi
- 5. Analisis Perilaku Pelanggan
  - a. Lakukan analisis lebih mendalam tentang karakteristik dan preferensi pelanggan di setiap kategori tarif
  - b. Gunakan wawasan ini untuk mengembangkan penawaran yang lebih disesuaikan dan strategi komunikasi yang terarah

#### **7.2.10 Lead Time Trend by Arrival Date Year**



#### **Evaluasi Terhadap Target**

Aspek	Detail
Nilai Aktual	124,66 atau 125 hari
Target	$\geq 100$ hari
Status	Grafik meningkat (Melebihi target)
Selisih dengan Target	+24 hari

### Kekuatan

1. **Lead time konsisten meningkat**, menunjukkan perencanaan dan pemesanan oleh pelanggan dilakukan lebih awal, yang memberikan lebih banyak waktu bagi manajemen untuk mengatur operasional.
2. **Nilai aktual melebihi target**, menandakan bahwa strategi marketing atau awareness program yang dijalankan telah mendorong pelanggan untuk memesan lebih awal.
3. **Peningkatan efisiensi operasional** dapat terjadi karena adanya buffer waktu lebih panjang antara pemesanan dan kedatangan tamu.

### Rekomendasi Strategis

1. **Pertahankan dan perluas program early booking** (misalnya diskon untuk pemesanan 3 bulan sebelumnya) untuk menjaga lead time tetap tinggi atau meningkat.
2. **Manfaatkan lead time panjang untuk perencanaan sumber daya** seperti pengadaan bahan makanan, penjadwalan staf, dan promosi khusus berdasarkan tren pemesanan.
3. **Analisis lebih lanjut** untuk mengetahui segmen pelanggan atau channel yang berkontribusi besar terhadap peningkatan lead time agar dapat difokuskan dalam strategi pemasaran berikutnya.

## **BAB VIII KESIMPULAN DAN SARAN**

### **8.1. Kesimpulan**

Penerapan sistem Data Warehouse dan Business Intelligence (DWBI) dilakukan untuk menganalisis data pemesanan hotel secara lebih efektif. Dengan memanfaatkan dataset Hotel Booking Demand dari Kaggle, analisis dilakukan terhadap perilaku pelanggan, pola pembatalan, serta karakteristik pemesanan. Hasil eksplorasi menunjukkan bahwa sekitar 37% reservasi dibatalkan, terutama yang memiliki lead time panjang. Berdasarkan temuan tersebut, dibangunlah struktur data warehouse berbasis star schema yang terdiri dari satu tabel fakta dan sejumlah tabel dimensi, lalu diimplementasikan dalam MySQL dengan proses ETL menggunakan Pentaho.

Implementasi data mining seperti CatBoost Classifier dan K-Means juga diterapkan untuk memprediksi pembatalan serta mengelompokkan pelanggan berdasarkan pola reservasi. Analisis tersebut kemudian divisualisasikan dalam dashboard interaktif dengan pendekatan Balanced Scorecard, yang mencakup indikator seperti durasi menginap, tingkat pembatalan, pemenuhan permintaan khusus, dan loyalitas tamu. Pendekatan ini memberikan manfaat nyata bagi manajemen hotel dalam pengambilan keputusan berbasis data, peningkatan efisiensi operasional, serta pemahaman yang lebih baik terhadap perilaku pelanggan.

## 8.2. Saran

Sistem yang dibangun dapat berjalan lebih optimal proses pengecekan dan pembersihan data sebaiknya dilakukan secara lebih menyeluruh. Meskipun data yang digunakan sudah cukup baik, tetapi ada beberapa bagian yang bisa diperbaiki untuk menghasilkan analisis yang lebih akurat dan minim kesalahan.

Tampilan dashboard dapat ditingkatkan dengan menambahkan elemen interaktif seperti filter waktu, jenis hotel, atau asal negara tamu. Hal ini bertujuan agar pengguna dapat menyesuaikan tampilan data sesuai kebutuhan dan lebih mudah memahami informasi yang disajikan.

Selain itu, pengembangan model analisis dapat diperluas dengan mencoba metode lain sebagai pembanding, untuk melihat kemungkinan hasil yang lebih baik atau temuan baru dari pola data yang ada. Akan lebih baik pula jika sistem dapat terhubung langsung dengan data hotel secara otomatis (real-time), sehingga dashboard selalu menampilkan informasi terbaru tanpa perlu pembaruan manual. Menyusun dokumentasi kerja secara rapi dan jelas, mulai dari tahap pengolahan data hingga pembuatan dashboard. Hal ini akan mempermudah tim lain jika ingin melakukan pengembangan lebih lanjut di masa mendatang.

## PENGERJAAN TUGAS ANGGOTA KELOMPOK

NAMA	NIM	PENGERJAAN
Ahmad Fauzi	1202220263	<p>Menulis laporan tugas besar.</p> <p><b>Pengerjaan:</b></p> <p><b>BAB 1:</b> Membuat Objektif, Balance Scorecard, dan Finalisasi daftar KPI.</p> <p><b>BAB 2:</b> Mencari dataset pada kaggle dan melakukan EDA.</p> <p><b>BAB 3:</b> Melakukan perancangan dan dokumentasi struktur Star Schema.</p> <p><b>BAB 4:</b> Merancang bentuk database, melakukan DDL, dan Dokumentasi Desain Struktur Database.</p> <p><b>BAB 5:</b> Melakukan ETL pada fact_booking, dim_guest, dim_date, dan dim_hotel.</p> <p><b>BAB 6:</b> Melakukan penggeraan data mining bagian klasifikasi menggunakan CatBoost.</p>

		<b>BAB 7:</b> Melakukan export database, Load database pada HeidiSQL agar bisa diakses secara online melalui aiven cloud, melakukan pembuatan dan perhitungan visualisasi dari KPI yang telah dibuat, dan melakukan analisis serta evaluasi pada masing-masing visualisasi sesuai dengan target KPI pada BAB 1.
Nerlis Fitria Nurani	1202223307	<p>Menulis laporan tugas besar</p> <p><b>Pengerjaan:</b></p> <p><b>Bab 1:</b> Membantu menentukan KPI</p> <p><b>Bab 3:</b> Membantu melakukan perancangan Star Schema</p> <p><b>Bab 5:</b> Melakukan ETL pada dim_country, dim_deposite-type, dim_reservation_status</p> <p><b>Bab 7:</b> Membantu perancangan dashboard visualisasi KPI, penjelasan KPI 8 serta membantu analisis dan evaluasi pada visualisasi sesuai dengan target KPI pada BAB 1</p>
Rafie Safaraz Aribowo	1202223025	<p>Menulis laporan tugas besar</p> <p><b>Pengerjaan:</b></p> <p><b>Bab 1:</b> Membantu menentukan KPI.</p> <p><b>Bab 3:</b> Membantu melakukan perancangan Star Schema.</p> <p><b>Bab 5:</b> Melakukan ETL pada dim_company dan dim_customer_type.</p> <p><b>Bab 7:</b> Membantu perancangan dashboard visualisasi KPI serta membantu analisis dan evaluasi pada masing-masing visualisasi sesuai dengan target KPI pada BAB 1.</p>
Ryannisa Syarifa Triandini	1202223163	<p>Menulis laporan tugas besar</p> <p><b>Pengerjaan:</b></p> <p><b>Bab 1:</b> Membantu menentukan KPI</p> <p><b>Bab 3:</b> Membantu melakukan perancangan Star Schema</p> <p><b>Bab 5:</b> Melakukan ETL pada dim_room, dim_meal,</p>

		<p>dim_market_segment dan penjelasan terkait dimension table</p> <p><b>Bab 7:</b> Membantu perancangan dashboard visualisasi KPI, penjelasan setiap KPI serta membantu analisis dan evaluasi pada masing-masing visualisasi sesuai dengan target KPI pada BAB 1</p>
Sarah Luki Raihani	1202223084	<p>Menulis laporan tugas besar</p> <p><b>Pengerjaan:</b></p> <p><b>Bab 1:</b> Membantu menentukan KPI</p> <p><b>Bab 3:</b> Membantu melakukan perancangan Star Schema</p> <p><b>Bab 5:</b> Melakukan ETL pada dim_distributin_channel dan dim_agent</p> <p><b>Bab 6:</b> Melakukan pengerojan data mining bagian klasterisasi menggunakan K-Means</p> <p><b>Bab 7:</b> Membantu perancangan dashboard visualisasi KPI serta membantu analisis dan evaluasi pada masing-masing visualisasi sesuai dengan target KPI pada BAB 1</p>

## **LAMPIRAN**

Link Dashboard:

<https://lookerstudio.google.com/s/v0GD5YzPvS8>

File ETL Pentaho (.ktr):

[https://drive.google.com/file/d/1yahYvdZsSPLDEKEMz\\_f1AkyIkQyNaPVi/view?usp=sharing](https://drive.google.com/file/d/1yahYvdZsSPLDEKEMz_f1AkyIkQyNaPVi/view?usp=sharing)

File Final Database (.sql):

[https://drive.google.com/file/d/1jGYcXa6QrsBTyJvI\\_fcG3NZQNbLy8gfi/view?usp=sharing](https://drive.google.com/file/d/1jGYcXa6QrsBTyJvI_fcG3NZQNbLy8gfi/view?usp=sharing)

Link GitHub (Dataset & Notebook (.ipynb)):

<https://github.com/ahmfzui/tubes-dwbi.git>