

Data Analytics for Business 2024

MID EXAM

Anggota Kelompok 3:

Nabila Putri Asy Syifa

KM-CS18165

Girinda Decalzgi Azade

KM-CS18270

Rendy Achmadiansyah Mukti

KM-CS18375

BAB I

Pendahuluan

1.1 Latar belakang masalah

City Hotel dan Resort Hotel memiliki beberapa indikasi masalah yang tercatat dalam dataset mereka mulai dari tanggal 1 Juli 2015 hingga 31 Agustus 2017. Kedua perusahaan hotel ini menghadapi beberapa masalah dalam proses reservasi dan pelayanan yang mempengaruhi tingkat kepuasan pelanggan hingga berdampak pada pembatalan yang sering terjadi dalam kurun waktu tersebut. Indikasi masalah-masalah tersebut dapat diidentifikasi dalam beberapa bagian sebagai berikut.

1.1.1 Tingkat Pembatalan dan Alur Reservasi

Perusahaan hotel mengalami masalah seperti tingginya tingkat pembatalan, variasi lead time yang tidak terprediksi, ketidakefektifan dalam menangani permintaan khusus, proses pembayaran yang tidak efektif, dan kurangnya komunikasi saat proses check-in dan check-out yang berdampak negatif pada pengalaman pelanggan dan efisiensi operasional. Melalui analisis dataset yang mencakup informasi pembatalan, lead time, permintaan khusus, *Average Daily Rate* dan status reservasi, dilakukan pemetaan alur reservasi dan pembuatan diagram proses bisnis (BPMN) untuk mengidentifikasi tahapan proses reservasi. Output dari analisis ini mencakup alur reservasi dari pemesanan hingga check-out, serta analisis tiap permasalahan seperti tingkat pembatalan dan variasi lead time yang divisualisasikan dalam grafik. Hasil ini diharapkan dapat memberikan wawasan bagi hotel dalam mengurangi jumlah pembatalan, meningkatkan kepuasan pelanggan, serta memperbaiki manajemen reservasi dan sumber daya untuk meningkatkan pendapatan.

1.1.2 Kualitas Dataset Reservasi Hotel

Pada masalah kualitas data dalam dataset reservasi hotel, yang mencakup data "kotor" yang dapat memengaruhi hasil analisis dan pengambilan keputusan. Data yang kotor ini sering kali muncul dalam bentuk duplikasi, inkonsistensi, dan data yang hilang, sehingga diperlukan proses pembersihan data (*data preprocessing*) untuk mengidentifikasi dan memperbaiki permasalahan tersebut. Duplikasi data dapat

mengacaukan perhitungan tingkat pembatalan atau permintaan reservasi, sementara inkonsistensi dalam format atau nilai data bisa mengarah pada kesimpulan yang salah mengenai pola permintaan atau kebiasaan pelanggan. Dalam tahap pembersihan ini, data akan disusun ulang untuk menghilangkan entri duplikat, mengoreksi inkonsistensi format, dan menangani data yang hilang melalui metode yang sesuai, seperti pengisian nilai atau penghapusan baris yang tidak relevan. Hasil dari proses ini adalah dataset bersih yang akan menjadi dasar bagi analisis lebih lanjut. Dataset yang bersih ini akan membantu memastikan bahwa analisis yang dilakukan berikutnya dapat memberikan hasil yang lebih akurat dan dapat mendukung pengambilan keputusan bisnis yang tepat.

1.1.3 Evaluasi Kinerja Hotel dan Pola Reservasi Pelanggan

Dilakukan analisis data pada database reservasi hotel menggunakan SQL yang berfokus pada evaluasi kinerja hotel dan pemahaman pola reservasi pelanggan. Beberapa metrik penting yang dianalisis meliputi tingkat pembatalan reservasi per tahun berdasarkan jenis hotel, rata-rata lead time untuk setiap kelompok permintaan khusus, tanggal dengan pendapatan tertinggi dari reservasi yang selesai, serta hubungan antara lead time dan keberhasilan pembayaran. Analisis ini akan memberikan wawasan mendalam mengenai aspek-aspek seperti pola pembatalan, waktu tunggu pelanggan sebelum check-in, pendapatan harian rata-rata, dan korelasi lead time dengan keberhasilan pembayaran. Hasil dari analisis ini adalah laporan yang mendetail mengenai tren dan pola dalam data reservasi hotel, yang akan menjadi dasar bagi keputusan untuk meningkatkan kinerja operasional dan pemanfaatan sumber daya hotel secara lebih efektif.

1.1.4 Data Exploratory Terhadap Dataset Reservasi Hotel

Analisis ini berfokus pada eksplorasi data secara mendalam (Exploratory Data Analysis atau EDA) untuk mengidentifikasi pola-pola penting dalam dataset reservasi hotel yang telah dibersihkan. Tahap ini menggunakan visualisasi grafik untuk membantu memahami hubungan antara variabel seperti rasio lead time terhadap pembatalan, perbedaan pola pendapatan antara Resort Hotel dan City Hotel, asal negara pelanggan yang sering membatalkan reservasi, serta perilaku pemesanan last-minute.

Melalui grafik-grafik ini, hotel dapat memperoleh wawasan tentang faktor-faktor yang memengaruhi pembatalan, pendapatan, dan perilaku pemesanan, sehingga dapat merancang strategi yang lebih sesuai dengan kebutuhan dan preferensi pelanggan. Hasil yang akan didapatkan adalah serangkaian visualisasi dan analisis yang memberikan gambaran menyeluruh tentang tren reservasi yang akan mendukung hotel dalam mengambil keputusan untuk mengoptimalkan tingkat pendapatan.

1.1.5 Perbedaan *Average Daily Rate* (ADR) antara City Hotel dan Resort Hotel

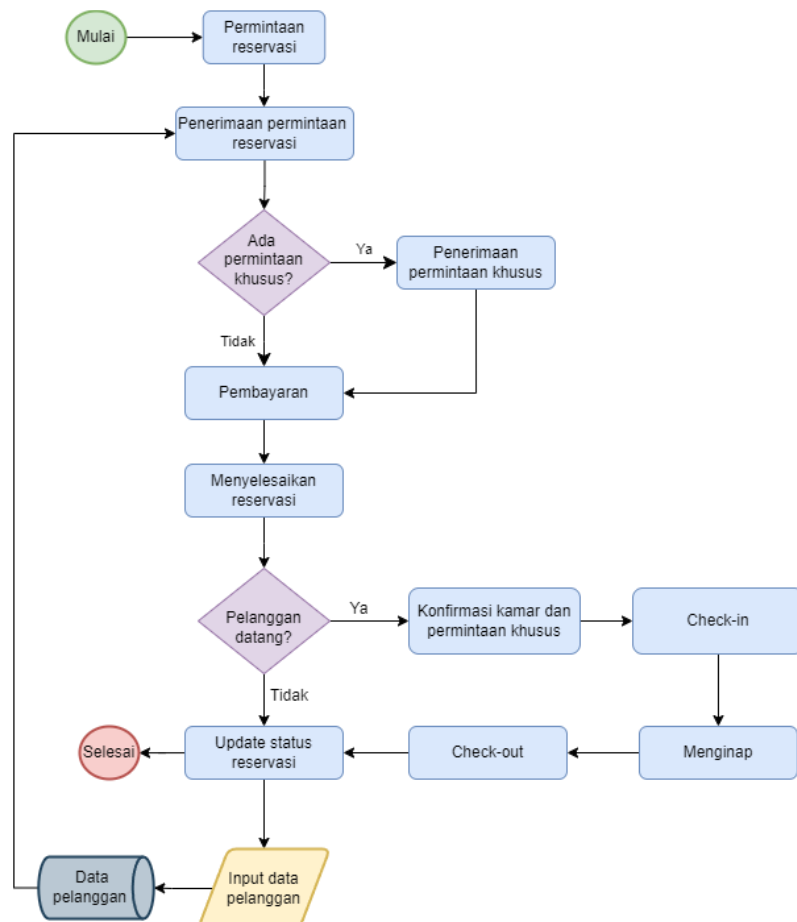
A/B testing yang dilakukan digunakan untuk menguji hipotesis mengenai perbedaan rata-rata tarif harian *Average Daily Rate* atau ADR antara City Hotel dan Resort Hotel. Hipotesis awal (H_0) menyatakan bahwa rata-rata ADR kedua jenis hotel sama, sedangkan hipotesis alternatif (H_1) menyatakan bahwa ada perbedaan antara keduanya. Dengan analisis statistik ini, hasil dari A/B testing akan menunjukkan apakah ada bukti yang cukup untuk menerima atau menolak hipotesis awal. Hasilnya adalah kesimpulan mengenai hipotesis mana yang diterima sehingga akan membantu hotel memahami apakah perbedaan jenis hotel memengaruhi tarif harian. Selain itu, dapat mendukung strategi penetapan harga yang lebih efektif.

BAB II

Business Process Analysis

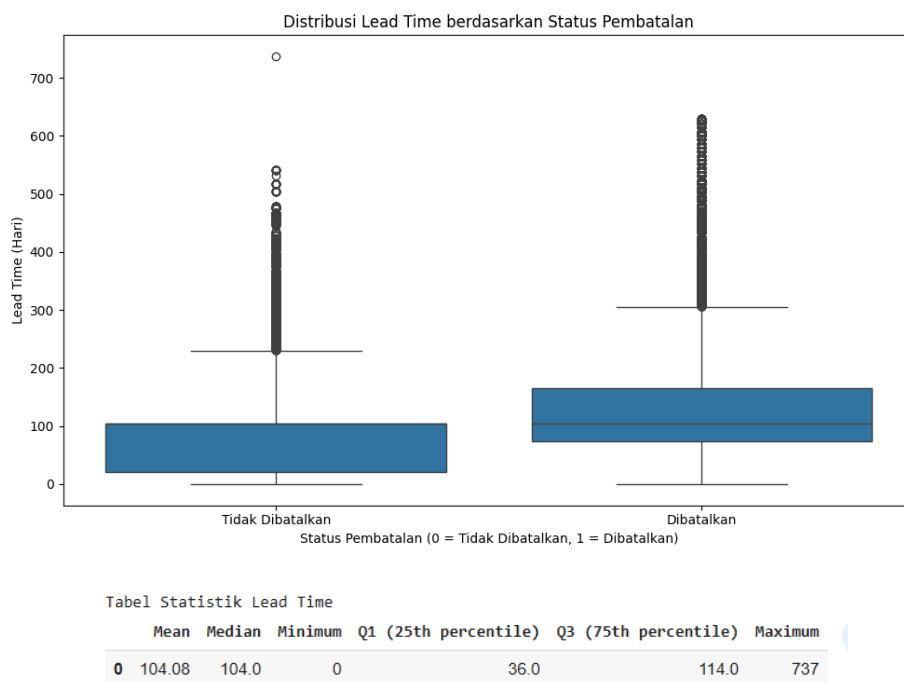
2.1 Identifikasi Masalah

Hotel menghadapi sejumlah masalah terkait tingkat pembatalan yang tinggi, ketidakpastian dalam variasi lead time, kesulitan dalam memenuhi permintaan khusus, proses pembayaran yang kurang efisien, serta komunikasi yang kurang baik selama check-in dan check-out. Semua isu ini berdampak buruk pada pengalaman pelanggan dan efisiensi operasional hotel. Oleh karena itu, analisis pertama yang dilakukan yaitu mengidentifikasi alur tahapan reservasi dari awal hingga akhir. Alur tahapan reservasi hotel disajikan dalam bentuk *flowchart* agar memudahkan dalam proses identifikasi masalah di setiap tahapannya. Berikut merupakan *flowchart* yang menggambarkan alur tahapan reservasi dapat dilihat pada Gambar 1.



Gambar 1. Flowchart Alur Tahapan Reservasi Hotel

Berdasarkan Gambar 1, reservasi hotel dimulai dengan adanya permintaan reservasi kemudian permintaan reservasi diterima. Pada tahap awal ini tidak terdapat pengecekan untuk ketersediaan kamar yang akan dipesan, karena berdasarkan kolom-kolom pada dataset tidak ada kolom yang menunjukkan konfirmasi terkait ketersediaan kamar. Hanya terdapat kolom *days_in_waiting_list* yang menunjukkan jumlah hari yang dipesan sedang berada dalam daftar antrian sebelum dikonfirmasi ke pelanggan. Informasi ini tidak cukup untuk digunakan karena tidak ada kolom yang menunjukkan kapan terjadinya konfirmasi tersebut. Kita juga tidak dapat menghitung berapa hari yang akhirnya berhasil dikonfirmasi ke pelanggan. Padahal variasi lead time yang tinggi memerlukan kepastian untuk ketersediaan kamar. Pelanggan yang melakukan reservasi jauh – jauh hari dapat menjadi prioritas untuk sering diberi konfirmasi terkait reservasinya agar menurunkan jumlah pembatalan. Pada Gambar 2 dan Gambar 3 berikut merupakan analisis terkait variasi lead time dan tingkat pembatalan.

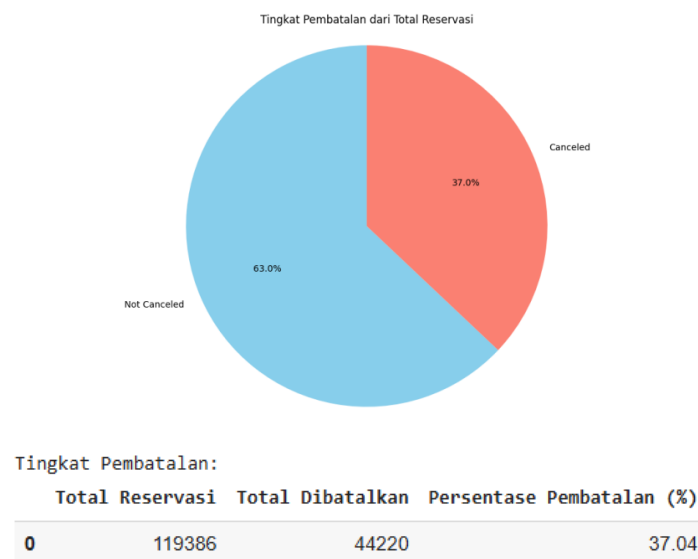


Gambar 2. Distribusi Lead Time

Tabel statistik lead time memberikan informasi penting seperti rata-rata, median, dan nilai kuartil. Berdasarkan tabel statistik, nilai mean lead time berada di 104,08 hari, sedangkan median titik tengah data juga di 104 hari. Ini menunjukkan bahwa sebagian besar data lead time terkonsentrasi di sekitar angka ini.

Boxplot yang ditampilkan menunjukkan distribusi lead time berdasarkan status pembatalan. Lead time untuk reservasi yang dibatalkan lebih tinggi dibandingkan yang tidak dibatalkan. Dalam boxplot untuk distribusi variasi lead time berdasarkan status pembatalan, outlier mulai muncul pada lead time di atas 200 hari hingga 737 hari (nilai maksimum). Namun, dengan nilai minimum yang sangat rendah, yaitu 0 hari, dan maksimum yang sangat tinggi, yaitu 737 hari, terdapat perbedaan yang mencolok dalam durasi lead time pelanggan. Lead time di atas 200 hari termasuk kategori outlier karena jauh di atas nilai tipikal data (di sekitar mean dan median). Outlier ini mencerminkan reservasi yang dilakukan sangat jauh sebelumnya, yang kemungkinan besar merupakan reservasi spesifik atau dari pelanggan dengan preferensi pemesanan awal yang ekstrem.



Adanya outlier ini menunjukkan ketidakpastian yang signifikan dalam lead time, yang dapat menyulitkan hotel dalam perencanaan kamar dan alokasi sumber daya. Hal ini bisa berdampak pada efisiensi operasional dan risiko overbooking atau underbooking yang lebih tinggi, tergantung pada kebutuhan pelanggan yang tidak biasa atau sulit diprediksi. Outlier lead time yang sangat panjang hingga 737 hari, meningkatkan risiko pembatalan tinggi karena pelanggan lebih mungkin berubah rencana dalam rentang waktu yang lama. Dampaknya, hotel menghadapi risiko kehilangan pendapatan karena pembatalan mendadak, terutama jika terjadi saat mendekati tanggal kedatangan. Perhatikan Gambar 3 untuk mengetahui tingginya tingkat pembatalan yang terjadi.

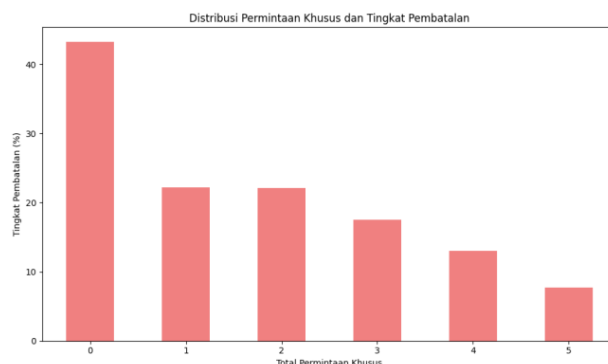


Gambar 3. Tingkat Pembatalan

Code yang dijalankan digunakan untuk menunjukkan tingginya jumlah pembatalan dari total reservasi yang ada. Pada tabel Tingkat Pembatalan, total reservasi berjumlah 119386, sedangkan total reservasi yang dibatalkan berjumlah 44220. Hal ini berarti bahwa tingkat pembatalan memang cukup tinggi yaitu mencapai sekitar 37% dari total reservasi. Pada visualisasi grafik yang menunjukkan perbandingan persentase untuk status pembatalan yaitu “Canceled” yang berarti dibatalkan dan “Not Canceled” untuk tidak dibatalkan dari total keseluruhan reservasi yang sudah terhitung pada tabel. Persentase “Not Canceled” yaitu 63% sedangkan persentase “Canceled” adalah 37%, artinya tingkat pembatalan cukup tinggi hingga lebih dari sepertiga dari total reservasi dan lebih dari setengahnya dari persentase “Not Canceled.” Pembatalan yang tinggi ini dapat mencerminkan masalah dalam pengalaman pelanggan, seperti kurangnya komunikasi, ketidakpuasan terhadap layanan, atau ketidakmampuan untuk memenuhi permintaan khusus. Maka, akan dilakukan analisis lebih lanjut mengenai hal-hal yang berhubungan dengan pengalaman pelanggan. Perhatikan Gambar 4 yang menampilkan hasil dari distribusi jumlah permintaan khusus dan tingkat pembatalannya.

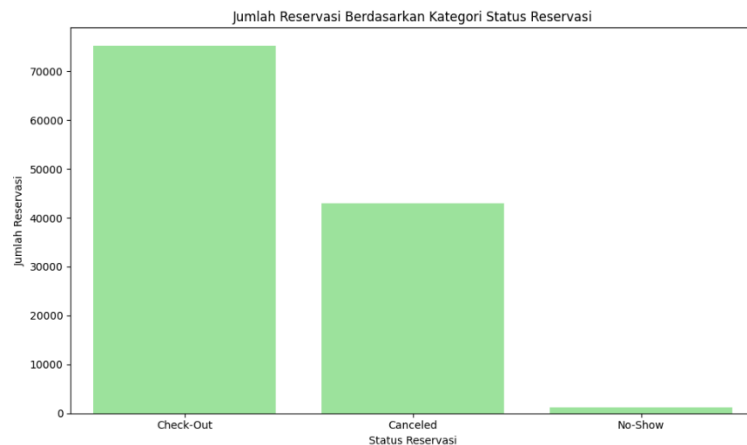
Distribusi Permintaan Khusus dan Tingkat Pembatalannya:

	Total Special Requests	Cancellation Rate (%)	
0	0	43.21	 
1	1	22.18	
2	2	22.06	
3	3	17.51	
4	4	12.99	
5	5	7.69	



Gambar 4. Distribusi Permintaan Khusus dan Tingkat Pembatalan

Berdasarkan Gambar 4 di atas, jumlah permintaan khusus berperan dalam tingginya jumlah pembatalan, dimana pelanggan yang tidak meminta permintaan khusus atau jumlah permintaan khususnya adalah 0, memiliki persentase pembatalan yang paling tinggi dan berjarak jauh dari pelanggan yang memiliki permintaan khusus mulai dari 1 hingga 5 permintaan khusus. Berdasarkan Gambar 4, semakin sedikit jumlah permintaan khusus dari pelanggan, maka semakin tinggi pula persentase pembatalannya, yaitu mencapai 43,21%. Begitu juga jika semakin banyak permintaan khususnya maka persentase pembatalannya semakin rendah yaitu bisa mencapai 7,69%. Jika permintaan khusus dapat dilayani dengan lebih baik lagi, maka perusahaan hotel dapat menekan angka pembatalan agar menjadi lebih rendah. Ini merupakan salah satu cara yang dapat dilakukan untuk mengatasi masalah tingginya tingkat pembatalan. Informasi mengenai permintaan khusus harus disampaikan dengan baik ke pelanggan baik dari proses reservasi, check-in, hingga check-out. Komunikasi yang buruk selama proses check-in dan check-out dapat mengakibatkan pengalaman pelanggan yang negatif. Ketika informasi tidak disampaikan dengan baik, pelanggan mungkin merasa bingung atau tidak puas, yang dapat berkontribusi pada jumlah pembatalan. Perhatikan Gambar 5 yang menunjukkan jumlah reservasi untuk setiap kategori status reservasi.



Tabel Jumlah Reservasi Berdasarkan Kategori Status Reservasi:

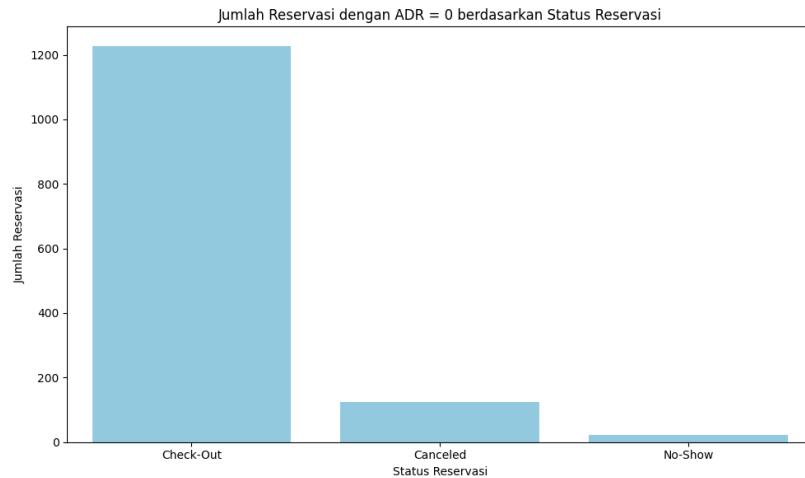
	Reservation Status	Count
0	Check-Out	75166
1	Canceled	43013
2	No-Show	1207

Gambar 5. Jumlah Reservasi Setiap Kategori Status Reservasi

Berdasarkan gambar 5 terlihat bahwa jumlah reservasi dengan status “Check-Out” adalah 75166, yang menunjukkan sebagian besar reservasi berhasil diselesaikan. Namun, jumlah yang cukup besar pada status “Canceled” (43013) dan “No-Show” (1207) yang menunjukkan potensi masalah komunikasi dengan pelanggan. Tingginya angka No-Show bisa menandakan bahwa pelanggan mungkin tidak menerima informasi atau pengingat yang memadai sebelum check-in, yang dapat menyebabkan ketidakhadiran. Demikian pula, jumlah pembatalan yang tinggi dapat menunjukkan ketidakpuasan terhadap proses reservasi atau kurangnya kejelasan mengenai kebijakan yang diberlakukan.

Langkah perbaikan dapat mencakup peningkatan sistem komunikasi, seperti pengiriman pengingat melalui email atau pesan teks terkait informasi reservasi, prosedur check-in, dan kebijakan pembatalan. Selain itu, hotel juga dapat mengumpulkan umpan balik dari pelanggan yang sering membatalkan atau tidak hadir untuk lebih memahami masalah yang perlu ditingkatkan dalam sistem reservasi mereka.

Komunikasi saat proses reservasi hingga check-out sangatlah penting, jika perusahaan hotel tidak segera melakukan penanganan maka angka pembatalan dapat semakin tinggi yang berdampak pada pendapatan hotel. Tetapi, masalah pendapatan hotel juga melibatkan layanan pembayaran pada hotel. Layanan pembayaran yang tidak efektif akan membuat pelanggan ragu untuk melanjutkan reservasi. Selain itu, Ketidaksesuaian informasi mengenai keberhasilan pembayaran yang dilakukan pelanggan dapat mengganggu pendapatan hotel. Pembayaran yang gagal ditandai dengan nilai ADR (Average Daily Rate) sebesar 0, menunjukkan adanya masalah dalam proses pembayaran. Ini bisa disebabkan oleh berbagai faktor, seperti kesalahan teknis, kebijakan pembayaran yang membingungkan, atau ketidakpuasan pelanggan terhadap opsi pembayaran yang tersedia. Perhatikan Gambar 6 untuk mengetahui lebih lanjut mengenai kegagalan pembayaran yang ditunjukkan berdasarkan kolom adr dan status reservasinya.



Tabel Jumlah Reservasi dengan ADR = 0 berdasarkan Status Reservasi:

	Reservation Status	Count
0	Check-Out	1226
1	Canceled	124
2	No-Show	21

Gambar 6. Jumlah Reservasi dengan ADR = 0 Berdasarkan Status Reservasi

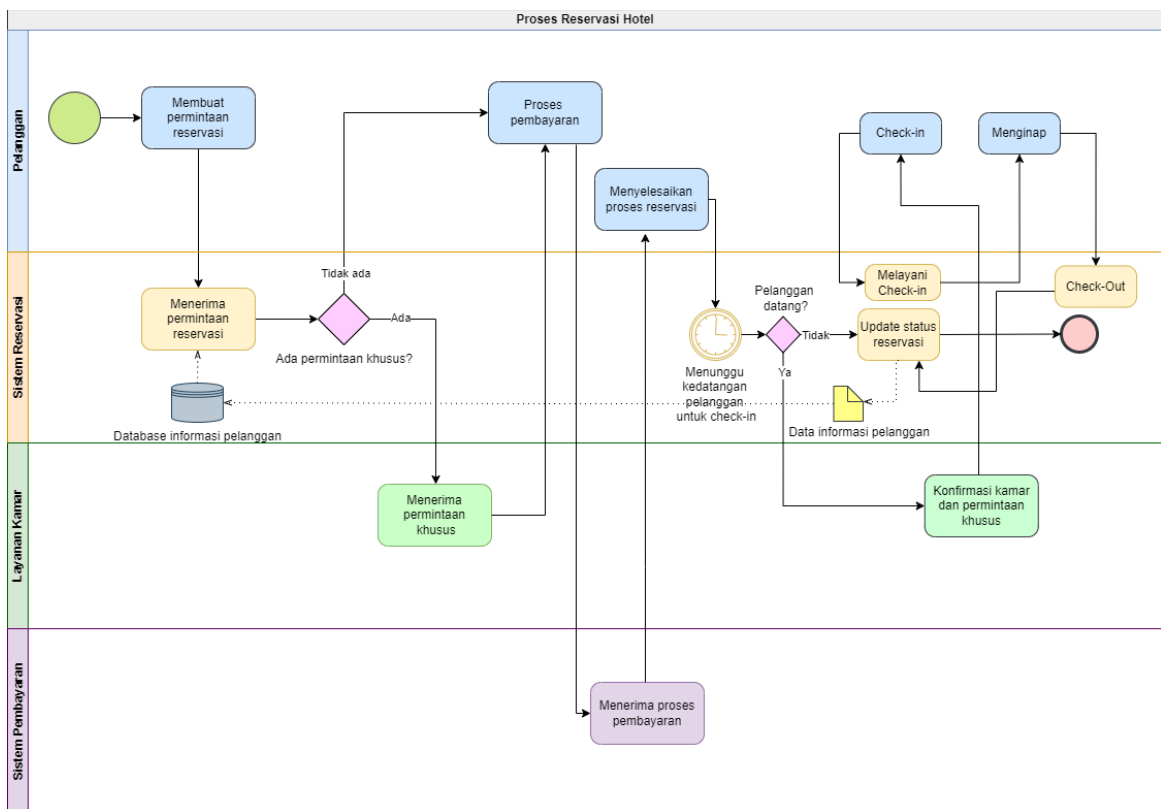
Berdasarkan Gambar 6, masalah $adr = 0$ saat status Check-Out yang berjumlah paling tinggi yaitu 1226 menunjukkan bahwa proses pembayaran yang tidak efisien memungkinkan tamu untuk menyelesaikan reservasi tanpa pembayaran yang tercatat, baik karena kendala dalam proses transaksi atau kegagalan verifikasi di sistem. Untuk mengatasi ini, hotel perlu memastikan sistem pembayaran untuk mengurangi risiko pembayaran gagal yang tidak tercatat. Verifikasi otomatis pada status reservasi agar `reservation_status` tidak bisa diubah menjadi "Check-Out" tanpa konfirmasi pembayaran yang sukses. Selain itu, perlu penyederhanaan proses pembayaran agar pelanggan merasa lebih yakin dalam menyelesaikan transaksi, meningkatkan kemungkinan mereka menyelesaikan pembayaran dan menurunkan pembatalan.

Perusahaan hotel perlu memperhatikan mengenai tahapan alur reservasi agar lebih efektif dan tidak menimbulkan berbagai masalah baru lainnya yang dapat meningkatkan tingkat pembatalan. Selain itu, peningkatan layanan mengenai permintaan khusus, komunikasi kepada pelanggan, dan pembayaran perlu segera ditangani. Solusi yang

diberikan dari hasil berbagai analisis ini dapat membantu hotel menurunkan angka pembatalan dan meningkatkan kepuasan pelanggan.

2.2 Diagram BPMN

Alur tahapan reservasi akan lebih jelas disajikan dalam diagram BPMN untuk mengetahui tindakan yang dilakukan oleh setiap aktor yang terlibat. Selain itu, akan lebih mudah untuk menemukan tahapan yang kurang optimal selama proses reservasi dari awal hingga akhir. Berikut diagram BPMN yang menyajikan alur proses reservasi hotel pada Gambar 7.



Gambar 7. Diagram BPMN Proses Reservasi Hotel

Dalam proses reservasi hotel ini, terdapat empat aktor utama yang terlibat yaitu pelanggan, sistem reservasi, layanan kamar, dan sistem pembayaran. Pelanggan memulai proses dengan mengajukan permintaan reservasi, melakukan pembayaran, melakukan check-in, hingga akhirnya check-out. Sistem reservasi berfungsi untuk mencatat permintaan pelanggan, termasuk memproses informasi terkait ketersediaan kamar dan status reservasi,

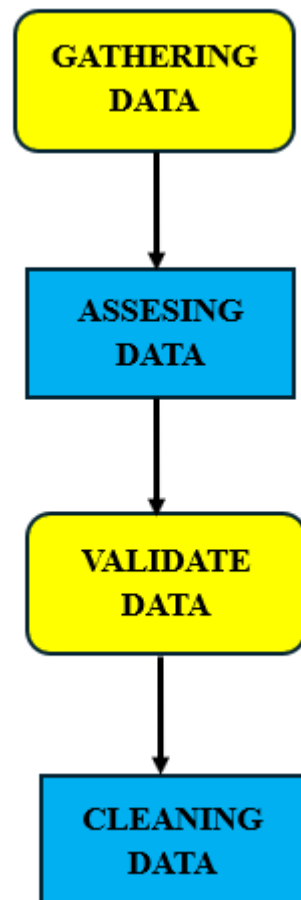
namun saat ini tidak ada mekanisme untuk mengecek ketersediaan kamar dan permintaan khusus secara otomatis, sehingga berpotensi menyebabkan masalah pada pemesanan. Hal ini menyebabkan pelanggan dapat memesan kamar yang mungkin tidak tersedia atau mengajukan permintaan khusus yang belum tentu bisa dipenuhi, sehingga berpotensi menurunkan kepuasan mereka dan meningkatkan pembatalan. Selain itu, tidak ada konfirmasi langsung dari sistem pembayaran terkait keberhasilan transaksi. Hal ini dapat menyebabkan pelanggan yang gagal membayar tetap melanjutkan proses reservasi, bahkan sampai tahap check-out, yang terlihat dari nilai ADR (Average Daily Rate) bernilai 0 pada reservasi yang sudah berstatus check-out. Untuk meningkatkan efektivitas, disarankan agar sistem reservasi dan sistem pembayaran harus melalui konfirmasi ketersediaan kamar, konfirmasi permintaan khusus, dan konfirmasi otomatis pembayaran yang berhasil agar proses reservasi lebih terstruktur dan akurat.

BAB III

Data Preprocessing and Structured Query Language

3.1 Data Pre-processing

Tahap pada bagian ini menjelaskan mengenai proses apa saja yang dilakukan dalam mempersiapkan data yang kotor untuk siap diolah atau dipakai. Dilihat pada gambar terdapat sebuah penggunaan flowchart sebagai penjelasan alur dalam tahap Data Pre-processing:



Gambar 8. Flowchart data processing yang dilakukan.

- **Gathering Data**

Tahap data pre-processing, dimulai dengan proses pengumpulan data (*gathering data*). Pada proses ini akan mengumpulkan semua data yang dibutuhkan untuk menjawab semua pertanyaan atau masalah bisnis yang ingin dihadapi. Disini dengan menggunakan Dataset **midterm_hotel_data.csv** dengan format CSV.

- **Assessing Data**

Setelah semua data yang dibutuhkan terkumpul, proses selanjutnya ialah penilaian terhadap data tersebut. Proses ini dilakukan untuk menilai kualitas dan struktur dari sebuah data. Selain itu, proses ini juga bertujuan untuk mengidentifikasi berbagai masalah yang terdapat dalam data, seperti missing value, duplicate data, inconsistent value, dan outlier.

- **Validate Data**

Setelah semua data sudah diidentifikasi atau di check dari berbagai masalah yang terdapat dalam data, seperti missing value, duplicate data, inconsistent value, dan outlier. Artinya, data sudah siap untuk di proses Cleaning Data.

- **Cleaning Data**

Apabila pada proses sebelumnya sudah menemukan masalah (missing value, duplicate data, inconsistent value, dan outlier) yang terdapat di dalam sebuah data, masalah tersebut harus dibersihkan sebelum masuk tahap analisis data. Dengan beberapa teknik yang dapat digunakan untuk membersihkan data, diantaranya:

1. **Teknik untuk Mengatasi Missing Value:**

- **Dropping**

Metode ini bekerja dengan cara menghapus seluruh baris atau kolom yang memiliki missing value. Didalam dataset **midterm_hotel_data.csv** terdapat atribut children yang dimana hanya memiliki missing value berjumlah 4 baris, maka solusi menanganinya adalah dengan melakukan dropping terhadap 4 baris yang terdapat missing value pada atribut children.

- **Imputation**

Metode ini bekerja dengan cara mengisi (*fill*) missing value dengan nilai tertentu. Pada data kontinu, bisa menggunakan nilai mean, median, atau mode sebagai pengganti missing value. Jika bekerja menggunakan data kategoris, dapat mengisi missing value dengan kategori yang paling sering muncul (modus).

Didalam dataset **midterm_hotel_data.csv** terdapat beberapa atribut yang memiliki missing value dengan jumlah yang sangat banyak karna hingga puluhan dan ratusan ribu row, yakni atribut `lead_time`, `stays_in_weekend_nights`, `adults`, `adr`, dan `total_of_special_requests` dimana solusi penanganan missing value dengan mengisinya secara mean, karena atribut tersebut termasuk kedalam type data kontinu. Sedangkan, untuk atribut `country`, `agent`, dan `company` dimana solusi penanganan missing value dengan mengisinya secara modus, karena atribut tersebut termasuk kedalam type data kategori.

2. Teknik Mengatasi Data Duplicate

Ketika menemukan duplikasi pada data, tentunya harus menghilangkan atau menghapus duplikasi tersebut. Karena didalam dataset **midterm_hotel_data.csv** tidak terdapat data duplicate jadi tidak perlu untuk penangana selanjutnya.

3. Teknik Mengatasi Data Inconsistent Value

Dengan solusi penangan ialah data type correction, yang artinya memastikan kolom memiliki type data yang benar, seperti yang terdapat pada atribut `reservation_status_date` dengan konversi dari object ke datetime, karena atribut tersebut berisi tentang tahun, bulan, dan tanggal. Lalu, untuk beberapa atribut type data numerik, seperti `lead_time`, `stays_in_weekend_nights`, `adults`, `children`, dan `total_of_special_requests` dengan konversi dari float ke integer, karena atribut tersebut untuk menghindari ketidaktepatan akibat angka desimal yang tidak relevan dan membantu menjaga konsistensi data di seluruh dataset. Selanjutnya, untuk beberapa atribut numerik yang di konversi menjadi type data kategori, seperti `country`, `agent`, dan `company`, karena atribut tersebut berisi tentang ID dan nama negara jadi itu cocok untuk type data untuk type data kategori sehingga membuat jelas makna data.

Dan berikutnya, melakukan drop kolom “Unnamed” karena hanya berisi nomor index saja, jadi data tersebut tidak diperlukan.

4. Teknik Mengatasi Data Outlier

Data yang memiliki nilai jauh di luar rentang normal dari dataset. Dimana pada atribut **adr** memiliki nilai jauh di luar rentang normal dataset **midterm_hotel_data.csv**. Maka dari itu, solusinya adalah dengan cara melakukan drop terhadap data yang outlier.

Setelah semua proses data pre-paration selesai, maka dataset **midterm_hotel_data.csv** sudah siap untuk dioleh sehingga nanti semua data yang ada bisa untuk menjawab semua pertanyaan atau masalah bisnis yang dihadapi.

Berikut adalah sebuah *source code* 1. python yang digunakan dalam proses data cleansing dari dataset **midterm_hotel_data.csv**.

Kode Program 1. Proses *data cleansing* pada Python.

```
1. # IMPORT LIBRARY
2. import pandas as pd
3. import numpy as np
4. import matplotlib.pyplot as plt
5. import seaborn as sns

6. # CONNECT TO GOOGLE DRIVE
7. from google.colab import drive
8. drive.mount('/content/drive')

9. # GATHERING DATA

10. # MEMBACA DATA FILE DAN MENYIMPAN KE DALAM DATAFRAME PANDAS
11. file_path = '/content/drive/MyDrive/dataset_mid_term/midterm_hotel_data.csv'
12. df = pd.read_csv(file_path)

13. # Menampilkan informasi struktur DataFrame
14. df.info()

15. # Menampilkan 3 baris pertama DataFrame
16. df.head(3)
17. # Menampilkan 5 baris pertama dari kolom ke-10 hingga ke-26
18. df.iloc[:, 10:27].head(5)
```

```
19. # ASSESSING DATA

20. # MENGHITUNG DAN MENAMPILKAN JUMLAH NILAI DATA "NULL"
21. df.isnull().sum()

22. # MENDETEKSI DAN MENGHITUNG JUMLAH DATA DUPLICATE
23. df.duplicated().sum()

24. # CLEANING DATA

25. # MENDETEKSI OUTLIER UNTUK ATRIBUT "ADR"
26. # mengecek outliers pada atribut adr menggunakan box plot
27. sns.boxplot(df['adr'])

28. # membuat histogram dengan seaborn untuk atribut adr
29. sns.histplot(df['adr'], bins=30)
30. plt.xlabel('Tarif Harian Rata-rata (ADR)')
31. plt.ylabel('Frekuensi')
32. plt.title('Histogram ADR')
33. plt.show()

34. # SOLUSI OUTLIER
35. # Tentukan Batas Outlier dengan IQR
36. Q1 = df['adr'].quantile(0.25)
37. Q3 = df['adr'].quantile(0.75)
38. IQR = Q3 - Q1
39. lower_bound = Q1 - 1.5 * IQR
40. upper_bound = Q3 + 1.5 * IQR

41. # Filter Dataframe untuk Menghapus Outlier
42. df_no_outliers = df[(df['adr'] >= lower_bound) & (df['adr'] <= upper_bound)]

43. # HASIL SETELAH DI BOUNDERY
44. # membuat histogram dengan seaborn untuk atribut adr
45. sns.histplot(df['adr'], bins=30)
46. plt.xlabel('Tarif Harian Rata-rata (ADR)')
47. plt.ylabel('Frekuensi')
48. plt.title('Histogram ADR')
49. plt.show()

50. # STATISTIK DEKRIPTIF DATAFRAME
51. df.describe()

52. # IMPUTATION & DATA NORMALIZATION PADA ATRIBUT "LEAD_TIME"
53. # Isi null value dengan mean pada atribut lead_time
54. df['lead_time'].fillna(df['lead_time'].mean(), inplace=True)
55. # mengubah format desimal pada atribut lead time menjadi format bilangan bulat
56. df['lead_time'] = df['lead_time'].astype(int)

57. # menampilkan informasi struktur DataFrame
58. df.info()

59. # menampilkan 5 baris pertama DataFram
60. df.head(5)

61. # menghitung dan menampilkan jumlah nilai data "NULL"
62. df.isnull().sum()

63. # IMPUTATION & DATA NORMALIZATION PADA ATRIBUT "STAYS_IN_WEEKEND_NIGHTS"
64. # Isi null value dengan mean pada atribut stays_in_weekend_nights
65. df['stays_in_weekend_nights'].fillna(df['stays_in_weekend_nights'].mean(), inplace=True)
66. # mengubah format desimal pada atribut stays_in_weekend_nights menjadi format bilangan bulat
67. df['stays_in_weekend_nights'] = df['stays_in_weekend_nights'].astype(int)
```

```
68. # menampilkan informasi struktur DataFrame
69. df.info()

70. # menampilkan 5 baris pertama DataFrame
71. df.head(5)

72. # menghitung dan menampilkan jumlah nilai data "NULL"
73. df.isnull().sum()

74. # IMPUTATION & DATA NORMALIZATION PADA ATRIBUT "ADULTS"
75. # Isi null value dengan mean pada atribut adults
76. df['adults'].fillna(df['adults'].mean(), inplace=True)
77. # mengubah format desimal pada atribut adults menjadi format bilangan bulat
78. df['adults'] = df['adults'].astype(int)

79. # menampilkan informasi struktur DataFrame
80. df.info()

81. # menampilkan kolom ke 10 sampai 26
82. df.iloc[:, 10:27].head(5)

83. # menghitung dan menampilkan jumlah nilai data "NULL"
84. df.isnull().sum()

85. # DROPPING & DATA NORMALIZATION PADA ATRIBUT "CHILDREN"
86. # Drop null value pada atribut children
87. df.dropna(subset=['children'], inplace=True)
88. # mengubah format desimal pada atribut children menjadi format bilangan bulat
89. df['children'] = df['children'].astype(int)

90. # menampilkan informasi struktur DataFrame
91. df.info()

92. # menampilkan kolom ke 10 sampai 26
93. df.iloc[:, 10:27].head(5)

94. # menghitung dan menampilkan jumlah nilai data "NULL"
95. df.isnull().sum()

96. # IMPUTATION PADA ATRIBUT "COUNTRY"
97. # Menampilkan Modus pada atribut country serta banyak jumlahnya
98. df['country'].value_counts()

99. # MENGISI JUMLAH BARIS DATA NULL DENGAN 3 VALUE MODUS SECARA ACAK
100. # Menghitung jumlah nilai NaN pada kolom 'country'
101. nan_count = df['country'].isna().sum()
102. # Membagi jumlah NaN untuk diisi dengan PRT, GBR, dan FRA secara hampir merata
103. prt_count = nan_count // 3
104. gbr_count = nan_count // 3
105. fra_count = nan_count - prt_count - gbr_count # Sisa nilai NaN untuk mengimbangi
106. # Membuat array yang berisi PRT, GBR, dan FRA dengan distribusi hampir merata
107. fill_values = np.array(['PRT'] * prt_count + ['GBR'] * gbr_count + ['FRA'] * fra_count)
108. # Mengacak nilai-nilai PRT, GBR, dan FRA agar diisi secara acak
109. np.random.shuffle(fill_values)
110. # Mengisi nilai NaN dengan nilai acak dari array fill_values
111. df.loc[df['country'].isna(), 'country'] = fill_values

112. # menghitung value mode
113. df['country'].value_counts()

114. # menampilkan kolom ke 10 sampai 26
115. df.iloc[:, 10:27].head(5)

116. # menghitung dan menampilkan jumlah nilai data "NULL"
117. df.isnull().sum()
```

```
118. # IMPUTATION & DATA NORMALIZATION PADA ATRIBUT "AGENT"
119. # Menampilkan Modus pada atribut agent serta banyak jumlahnya
120. df['agent'].value_counts()

121. # MENGISI JUMLAH BARIS DATA NULL DENGAN 2 VALUE MODUS SECARA ACAK
122. # Menghitung jumlah nilai NaN pada kolom yang ingin diisi
123. nan_count = df['agent'].isna().sum()
124. # Membagi jumlah NaN untuk diisi dengan 9.0 dan 240.0 secara seimbang
125. half_nan_count = nan_count // 2
126. # Membuat array yang menggabungkan 9.0 dan 240.0 dengan distribusi yang hampir merata
127. fill_values = np.array([9.0] * half_nan_count + [240.0] * (nan_count - half_nan_count))
128. # Mengacak nilai-nilai 9.0 dan 240.0 agar diisi secara acak
129. np.random.shuffle(fill_values)
130. # Mengisi nilai NaN dengan nilai acak dari array fill_values
131. df.loc[df['agent'].isna(), 'agent'] = fill_values
132. # Konversi kolom 'agent' ke tipe data string
133. df['agent'] = df['agent'].astype(str)
134. # Hilangkan belakang desimal koma
135. df['agent'] = df['agent'].str.replace('.0', '', regex=False)

136. # menghitung value mode
137. df['agent'].value_counts()

138. # menampilkan kolom ke 10 sampai 26
139. df.iloc[:, 10:27].head(5)

140. # menampilkan informasi struktur DataFrame
141. df.info()

142. # menghitung dan menampilkan jumlah nilai data "NULL"
143. df.isnull().sum()

144. # IMPUTATION & DATA NORMALIZATION PADA ATRIBUT "COMPANY"
145. # Menampilkan Modus pada atribut company serta banyak jumlahnya
146. df['company'].value_counts()

147. # MENGISI JUMLAH BARIS DATA NULL DENGAN 2 VALUE MODUS SECARA ACAK
148. # Menghitung jumlah nilai NaN pada kolom yang ingin diisi
149. nan_count = df['company'].isna().sum()
150. # Membagi jumlah NaN untuk diisi dengan 40.0 dan 223.0 secara seimbang
151. half_nan_count = nan_count // 2
152. # Membuat array yang menggabungkan 40.0 dan 223.0 dengan distribusi yang hampir merata
153. fill_values = np.array([40.0] * half_nan_count + [223.0] * (nan_count - half_nan_count))
154. # Mengacak nilai-nilai 40.0 dan 223.0 agar diisi secara acak
155. np.random.shuffle(fill_values)
156. # Mengisi nilai NaN dengan nilai acak dari array fill_values
157. df.loc[df['company'].isna(), 'company'] = fill_values
158. # Konversi kolom 'company' ke tipe data string
159. df['company'] = df['company'].astype(str)
160. # Hilangkan belakang desimal koma
161. df['company'] = df['company'].str.replace('.0', '', regex=False)

162. # menghitung value mode
163. df['company'].value_counts()

164. # menampilkan kolom ke 10 sampai 26
165. df.iloc[:, 10:27].head(5)

166. # menampilkan informasi struktur DataFrame
167. df.info()

168. # menghitung dan menampilkan jumlah nilai data "NULL"
169. df.isnull().sum()

170. # IMPUTATION PADA ATRIBUT "ADR"
171. # Isi null value dengan mean pada atribut adr
```

```
172. df['adr'].fillna(df['adr'].mean(), inplace=True)
173. # Format desimal hanya 1 angka setelah koma pada atribut adr
174. pd.options.display.float_format = '{:.1f}'.format

175. # menghitung dan menampilkan jumlah nilai data "NULL"
176. df.isnull().sum()

177. # IMPUTATION & DATA NORMALIZATION PADA ATRIBUT "TOTAL_OF_SPECIAL_REQUESTS"
178. # Isi null value dengan mean pada atribut total_of_special_requests
179. df['total_of_special_requests'].fillna(df['total_of_special_requests'].mean(), inplace=True)
180. # mengubah format desimal pada atribut total_of_special_requests menjadi format bilangan bulat
181. df['total_of_special_requests'] = df['total_of_special_requests'].astype(int)
182.
183. # menampilkan informasi struktur DataFrame
184. df.info()

185. # menampilkan 5 baris pertama DataFrame
186. df.head(5)

187. # menghitung dan menampilkan jumlah nilai data "NULL"
188. df.isnull().sum()

189. # DATA NORMALIZATION PADA ATRIBUT "RESERVATION_STATUS_DATE"
190. # Mengubah format object pada atribut reservation_status_date menjadi yang berkaitan dengan waktu
191. df['reservation_status_date'] = pd.to_datetime(df['reservation_status_date'])

192. # menampilkan informasi struktur DataFrame
193. df.info()

194. # menghitung dan menampilkan jumlah nilai data "NULL"
195. df.isnull().sum()

196. # DROP COLUMN 'Unnamed'
197. df.drop('Unnamed: 0', axis=1, inplace=True)

198. # MENYIMPAN DATAFRAME PANDAS KE DALAM FILE "CSV"
199. # download file csv hasil cleaning
200. df.to_csv('midterm_hotel_data.csv', index=False)
```

3.2 Data Extraction

Data Extraction ini bertujuan untuk menganalisis dan mengevaluasi berbagai metrik kinerja hotel, dengan fokus mengevaluasi kinerja hotel dan memahami pola reservasi pelanggan. Tahap ini mencakup analisis tingkat pembatalan reservasi, rata-rata waktu tunggu, tren pendapatan berdasarkan tanggal reservasi, serta hubungan antara lead time dan tingkat keberhasilan pembayaran.

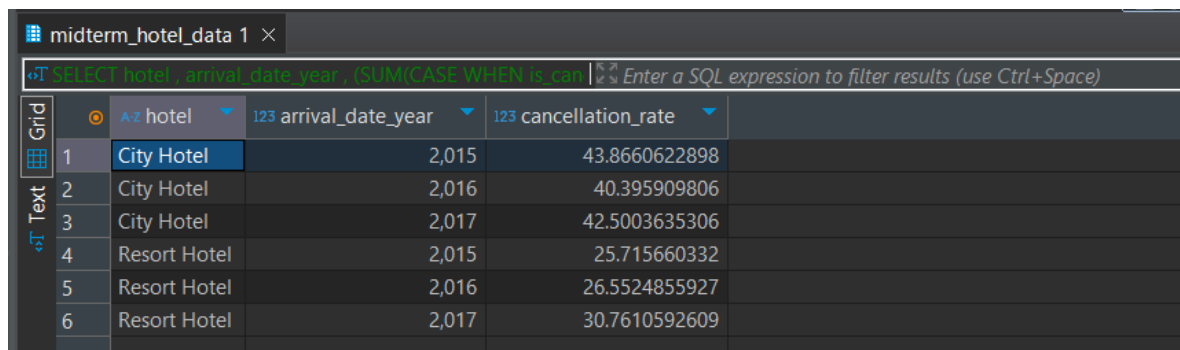
3.2.1 Tingkat Pembatalan Reservasi

Menghitung tingkat pembatalan reservasi untuk setiap jenis hotel berdasarkan tahun, bisa dilihat pada potongan Kode Program 1 berikut:

Kode Program 2. Menghitung tingkat pembatalan reservasi untuk setiap jenis hotel berdasarkan tahun.

```
1. SELECT
2.     hotel ,
3.     arrival_date_year ,
4.     (SUM(CASE WHEN is_canceled = 1 THEN 1 ELSE 0 END) * 100.0 /
5.      COUNT(*)) AS cancellation_rate
6. FROM   midterm_hotel_data
7. GROUP BY
8.     hotel, arrival_date_year;
```

Output dari Kode Program 1 menghasilkan tabel yang menunjukkan tingkat pembatalan reservasi berdasarkan jenis hotel dan tahun, seperti yang ditampilkan pada Gambar 3.2.1 Data ini menggambarkan persentase pembatalan pemesanan untuk dua jenis hotel: City Hotel dan Resort Hotel, dari tahun 2015 hingga 2017.



	hotel	arrival_date_year	cancellation_rate
1	City Hotel	2,015	43.8660622898
2	City Hotel	2,016	40.395909806
3	City Hotel	2,017	42.5003635306
4	Resort Hotel	2,015	25.715660332
5	Resort Hotel	2,016	26.5524855927
6	Resort Hotel	2,017	30.7610592609

Gambar 3.2.1 Output Kode Program 1.

Pada gambar 3.2.1 dapat dideskripsikan sebagai berikut:

1. Pada tahun 2015, hotel dengan tipe “City Hotel” memiliki presentase pembatalan pemesanan sebesar 43.8%. Sedangkan pada tipe “Resort Hotel”, memiliki presentase pembatalan pemesanan sebesar 25.7%.

2. Pada tahun 2016, hotel dengan tipe “City Hotel” memiliki presentase pembatalan pemesanan sebesar 40.3%. Sedangkan pada tipe “Resort Hotel”, memiliki presentase pembatalan pemesanan sebesar 26.5%.
3. Pada tahun 2017, hotel dengan tipe “City Hotel” memiliki presentase pembatalan pemesanan sebesar 42.5%. Sedangkan pada tipe “Resort Hotel”, memiliki presentase pembatalan pemesanan sebesar 30.7%.

Dari data ini, terlihat bahwa “City Hotel” cenderung memiliki tingkat pembatalan yang lebih tinggi dibandingkan dengan “Resort Hotel” pada setiap tahunnya. Selain itu, terjadi sedikit perubahan dari tahun ke tahun, yang menunjukkan adanya perbedaan pada pola pembatalan pelanggan untuk masing-masing jenis hotel.

3.2.2 Rata-Rata Waktu Tunggu

Untuk menilai rata-rata waktu tunggu (lead time) untuk reservasi dengan permintaan khusus, bisa dilihat pada potongan Kode Program 2 berikut:

Kode Program 3. Menghitung rata-rata waktu tunggu untuk reservasi dengan permintaan khusus di atas rata-rata lead time keseluruhan.

```
1. WITH avg_lead_time AS (  
2.     SELECT AVG(lead_time) AS overall_avg_lead_time  
3.     FROM midterm_hotel_data  
4. )  
5. SELECT  
6.     total_of_special_requests,  
7.     AVG(lead_time) AS average_lead_time  
8. FROM  
9.     midterm_hotel_data  
10. WHERE  
11.     lead_time > (SELECT overall_avg_lead_time FROM avg_lead_time)  
12. GROUP BY  
13.     total_of_special_requests;
```

Output dari Kode Program 2 menghasilkan tabel yang menunjukkan rata-rata lead time untuk setiap jumlah permintaan khusus yang memiliki lead time lebih tinggi dari rata-rata keseluruhan. Data ini membantu mengidentifikasi pola permintaan khusus berdasarkan lead time, yang memberikan gambaran waktu persiapan yang diinginkan pelanggan.

Enter a SQL expression to filter results (use Ctrl+Space)			
Grid	123 total_of_special_requests	123 average_lead_time	
1	0	220.7755460708	
2	1	199.6046725533	
3	2	190.6247160382	
4	3	188.0177383592	
5	4	201.4202898551	
6	5	171.5454545455	

Gambar 3.2.2 Output Kode Program 2.

Pada Gambar 3.2.2 data dapat dijelaskan sebagai berikut:

1. Reservasi tanpa permintaan khusus menunjukkan rata-rata lead time sebesar 220 hari.
2. Reservasi dengan 1 permintaan khusus menunjukkan rata-rata lead time sebesar 199 hari.
3. Reservasi dengan 2 permintaan khusus menunjukkan rata-rata lead time sebesar 190 hari.
4. Reservasi dengan 3 permintaan khusus menunjukkan rata-rata lead time sebesar 188 hari.
5. Reservasi dengan 4 permintaan khusus menunjukkan rata-rata lead time sebesar 201 hari.
6. Reservasi dengan 5 permintaan khusus menunjukkan rata-rata lead time sebesar 171 hari.

Dari data ini, terlihat bahwa rata-rata lead time cenderung lebih pendek pada reservasi dengan jumlah permintaan khusus yang lebih tinggi, meskipun terdapat sedikit variasi. Hal ini menunjukkan bahwa pelanggan yang memesan jauh hari cenderung tidak memerlukan banyak permintaan khusus, sementara pelanggan yang memesan lebih dekat dengan tanggal kedatangan lebih mungkin membuat permintaan khusus.

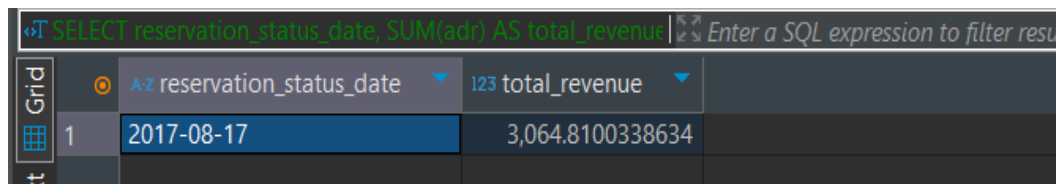
3.2.3 Menampilkan Tanggal Dengan Pendapatan Tertinggi

Mengidentifikasi tanggal reservasi dengan pendapatan tertinggi, bisa dilihat pada potongan Kode Program 3 berikut:

Kode Program 4. Menampilkan tanggal dengan pendapatan tertinggi.

```
1. SELECT
2.     reservation_status_date,
3.     SUM(adr) AS total_revenue
4. FROM
5.     midterm_hotel_data
6. WHERE
7.     total_of_special_requests > 2
8.     AND reservation_status = 'Check-Out'
9. GROUP BY
10.    reservation_status_date
11. ORDER BY
12.    total_revenue DESC
13. LIMIT 1;
```

Output dari kode program 3 menghasilkan tabel (bisa dilihat pada Gambar 3.2.3 yang menampilkan tanggal dengan pendapatan tertinggi dengan status pemesanan selesai atau “Check Out” dan juga pemesanan yang memiliki permintaan khusus lebih dari 2.



	reservation_status_date	total_revenue
1	2017-08-17	3,064.8100338634

Gambar 3.2.3 Output Kode Program 3.

Pada gambar 3.2.3 data dapat dideskripsikan bahwa tanggal dengan pendapatan tertinggi adalah tanggal 17 Agustus 2017 dengan total pendapatan sebanyak 3.064,81. Ini menunjukkan bahwa tanggal ini adalah salah satu yang memiliki tingkat pemesanan tinggi yang disertai permintaan layanan tambahan.

Dari data tersebut, dapat memberikan wawasan bagi pengelola hotel untuk mengenali periode dengan pendapatan tertinggi serta pola permintaan tambahan yang berpotensi meningkatkan kepuasan pelanggan dan meningkatkan pendapatan.

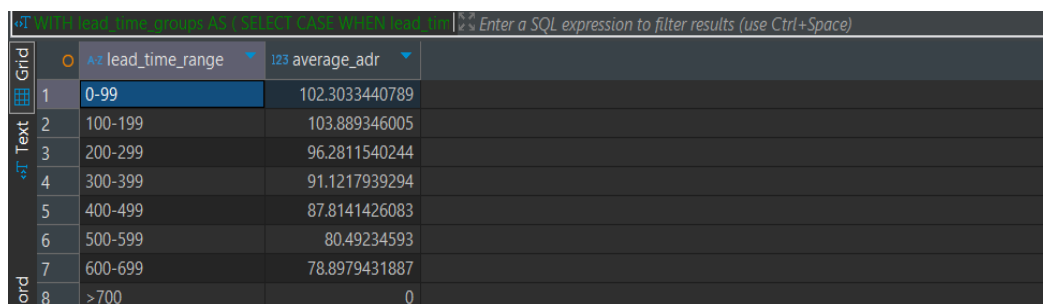
3.2.4 Hubungan antara Lead Time dan Keberhasilan Pembayaran

Untuk mengeksplorasi hubungan antara lead time dan average daily rate (ADR), bisa dilihat pada potongan Kode Program 4 berikut:

Kode Program 5. Hubungan antara Lead Time dan Keberhasilan Pembayaran.

```
1. WITH lead_time_groups AS (
2.     SELECT
3.         CASE
4.             WHEN lead_time BETWEEN 0 AND 99 THEN '0-99'
5.             WHEN lead_time BETWEEN 100 AND 199 THEN '100-199'
6.             WHEN lead_time BETWEEN 200 AND 299 THEN '200-299'
7.             WHEN lead_time BETWEEN 300 AND 399 THEN '300-399'
8.             WHEN lead_time BETWEEN 400 AND 499 THEN '400-499'
9.             WHEN lead_time BETWEEN 500 AND 599 THEN '500-599'
10.            WHEN lead_time BETWEEN 600 AND 699 THEN '600-699'
11.            WHEN lead_time > 700 THEN '>700'
12.            ELSE 'Unknown'
13.        END AS lead_time_range,
14.        adr
15.    FROM
16.        midterm_hotel_data
17. )
18. SELECT
19.     lead_time_range,
20.     AVG(adr) AS average_adr
21. FROM
22.     lead_time_groups
23. GROUP BY
24.     lead_time_range
25. ORDER BY
26.     lead_time_range;
```

Output dari Kode Program 4, yang ditampilkan pada Gambar 3.2.4 menunjukkan hubungan antara lead time dan keberhasilan pembayaran berdasarkan nilai rata-rata harian (ADR).



	lead_time_range	average_adr
1	0-99	102.3033440789
2	100-199	103.889346005
3	200-299	96.2811540244
4	300-399	91.1217939294
5	400-499	87.8141426083
6	500-599	80.49234593
7	600-699	78.8979431887
8	>700	0

Gambar 3.2.4 Output Kode Program 4

Pada Gambar 3.2.4 dapat dideskripsikan hubungan antara Lead Time dan keberhasilan pembayaran, sebagai berikut:

1. Lead time dalam rentang 0-99 hari menghasilkan rata-rata pembayaran sebesar 102,30.
2. Lead time dalam rentang 100-199 hari menghasilkan rata-rata pembayaran sebesar 103,88.
3. Lead time dalam rentang 200-299 hari menghasilkan rata-rata pembayaran sebesar 96,28.
4. Lead time dalam rentang 300-399 hari menghasilkan rata-rata pembayaran sebesar 91,12.
5. Lead time dalam rentang 400-499 hari menghasilkan rata-rata pembayaran sebesar 87,81.
6. Lead time dalam rentang 500-599 hari menghasilkan rata-rata pembayaran sebesar 80,49.
7. Lead time dalam rentang 600-699 hari menghasilkan rata-rata pembayaran sebesar 78,89.
8. Lead time lebih dari 700 hari tidak mencatat adanya pembayaran yang berhasil.

Dari data tersebut dapat disimpulkan bahwa semakin panjang lead time, rata-rata pembayaran cenderung menurun. Hal ini menunjukkan bahwa pelanggan yang melakukan reservasi jauh hari tampaknya memiliki kecenderungan untuk membatalkan atau menunda pembayaran. Sebaliknya, pelanggan yang melakukan reservasi dalam jangka waktu lebih singkat cenderung lebih pasti dalam pembayaran, dengan nilai ADR yang lebih tinggi. Analisis ini dapat membantu pengelola hotel dalam merancang strategi pembayaran dan kebijakan pembatalan sesuai dengan pola lead time pelanggan.

BAB IV

Python Programming

4.1 Exploratory Data Analysis

Pada tahap ini, bertujuan untuk melakukan proses Exploratory Data Analysis dari hasil dataset **midterm_hotel_data.csv** yang sebelumnya sudah di cleaning datanya, karena dengan dataset yang sudah di cleaning membuat informasi dalam dataset menjadi valid. Dengan proses Exploratory Data Analysis dapat memberikan kejelasan terhadap struktur data dan memberikan wawasan yang mendalam tentang data dan membantu dalam pengambilan keputusan yang lebih baik. Berikut adalah beberapa pertanyaan insight terhadap data hotel, diantaranya:

I. Bagaimana rasio antara **lead_time** dengan jumlah hari reservasi dibatalkan (**is_canceled == 1**)?

Kode Program 6. Proses *Exploratory Data Analysis* pada Python.

```
1. # Langkah 1: Filter data reservasi yang dibatalkan
2. canceled_reservations = df[df['is_canceled'] == 1]

3. # Langkah 2: Hitung total lead_time untuk reservasi yang dibatalkan
4. total_lead_time = canceled_reservations['lead_time'].sum()

5. # Langkah 3: Hitung total jumlah hari reservasi yang dibatalkan
6. total_days_canceled = (canceled_reservations['stays_in_weekend_nights'] +
7. canceled_reservations['stays_in_week_nights']).sum()

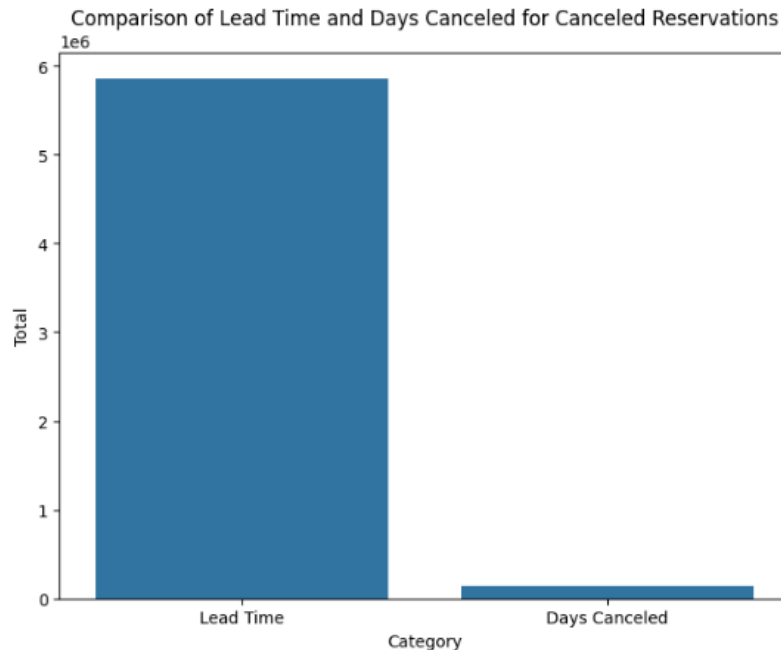
8. # Langkah 4: Hitung rasio antara lead_time dan jumlah hari reservasi yang dibatalkan
9. cancellation_ratio = total_lead_time / total_days_canceled
10. print(f"Rasio antara lead_time dan jumlah hari reservasi yang dibatalkan: {cancellation_ratio:.2f}")

11. # Data untuk visualisasi
    data = {'Category': ['Lead Time', 'Days Canceled'], 'Total': [total_lead_time, total_days_canceled]}

12. # Membuat bar chart
13. plt.figure(figsize=(8, 6))
14. sns.barplot(x='Category', y='Total', data=data)
15. plt.title('Comparison of Lead Time and Days Canceled for Canceled Reservations')
16. plt.ylabel('Total')
17. plt.show()
```

Output:

Rasio antara **lead_time** dan jumlah hari reservasi yang dibatalkan: 41.37



Gambar 4.1. Output Kode Program 6.

Wawasan yang di dapatkan:

Berdasarkan dari visualisasi diatas menampilkan bar “lead time” yang jauh lebih tinggi daripada bar “days canceled” yang secara visual mengkonfirmasi rasio yang tinggi. Artinya, secara keseluruhan rasio 41.37 ini menunjukkan bahwa rata-rata lead_time untuk reservasi yang dibatalkan adalah 41.37 hari untuk setiap hari yang di pesan dalam reservasi tersebut. Sehingga, mencerminkan kecenderungan pelanggan yang memesan jauh sebelum tanggal check-in, namun akhirnya membatalkan reservasi tersebut. Ini memberikan wawasan kepada hotel tentang kebutuhan untuk memahami ketidakpastian dalam rencana pelanggan dan mengembangkan strategi yang dapat membantu untuk mempertahankan reservasi, mengelola pembatalan dengan lebih baik, dan memaksimalkan pendapatan dari kamar yang dibatalkan.

Dengan memahami pola ini, hotel bisa merancang kebijakan yang lebih adaptif, baik dalam kebijakan pembatalan maupun dalam strategi promosi, guna mempertahankan tingkat hunian dan pendapatan yang optimal.

II. Apakah ada perbedaan pola pendapatan antara Resort Hotel dan City Hotel selama beberapa bulan tertentu?

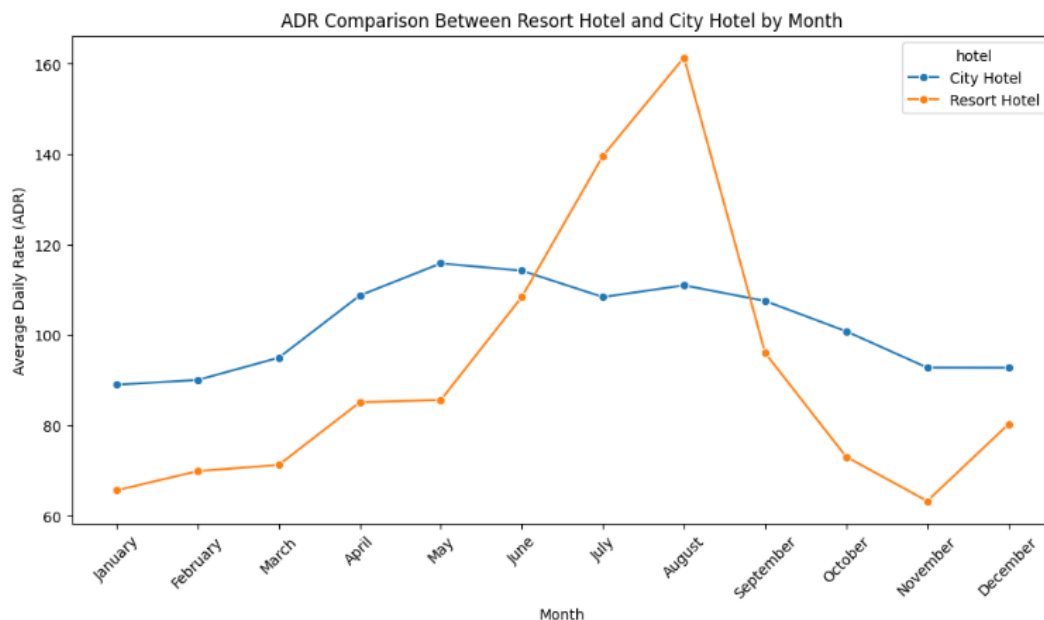
Kode Program 7. Proses *Exploratory Data Analysis* pada Python

```
1. # Group data by hotel type and month, calculating the average ADR (Average Daily Rate)
2. hotel_adr = df.groupby(['hotel', 'arrival_date_month'])['adr'].mean().reset_index()

3. # Sort the months in order (if needed)
4. months_order = ['January', 'February', 'March', 'April', 'May', 'June', 'July', 'August',
                    'September', 'October', 'November', 'December']
5. hotel_adr['arrival_date_month'] = pd.Categorical(hotel_adr['arrival_date_month'],
                                                    categories=months_order, ordered=True)

6. # Plotting the ADR trends for Resort Hotel and City Hotel
7. plt.figure(figsize=(12, 6))
8. sns.lineplot(data=hotel_adr, x='arrival_date_month', y='adr', hue='hotel', marker='o')
9. plt.title('ADR Comparison Between Resort Hotel and City Hotel by Month')
10. plt.xlabel('Month')
11. plt.ylabel('Average Daily Rate (ADR)')
12. plt.xticks(rotation=45)
13. plt.show()
```

Output:



Gambar 4.1. Output Kode Program 7.

Wawasan yang di dapatkan:

Berdasarkan dari analisis line plot diatas, dapat disimpulkan bahwa terdapat perbedaan pola pendapatan antara Resort Hotel dan City Hotel selama beberapa bulan tertentu. Karena, Perbedaan pola pendapatan ini dipengaruhi oleh faktor musiman. Sehingga, Resort Hotel memiliki pendapatan (ADR) yang lebih tinggi selama bulan-bulan musim panas, terutama Juli dan Agustus. Hal ini menunjukkan bahwa Resort Hotel lebih populer sebagai destinasi liburan musim panas. Sedangkan City Hotel memiliki pendapatan (ADR) yang lebih tinggi selama bulan-bulan musim semi (April-Mei) dan musim gugur (September-Oktober). City Hotel juga cenderung memiliki pendapatan yang lebih stabil sepanjang tahun dibandingkan dengan Resort Hotel. Ini juga terjadi karena di pengaruhi oleh faktor lainnya seperti lokasi, fasilitas, dan target pasar juga dapat mempengaruhi pola pendapatan.

Dengan pemahaman yang mendalam tentang wawasan ini memungkinkan Resort Hotel dan City Hotel untuk mengoptimalkan strategi mereka, memaksimalkan pendapatan, dan meningkatkan profitabilitas. Penting untuk diingat bahwa faktor-faktor yang mempengaruhi pola pendapatan dapat bervariasi dan perlu dipantau secara berkala untuk menyesuaikan strategi bisnis.

III. Turis dari negara mana sajakah yang sering melakukan cancelling reservation baik untuk Resort Hotel maupun City Hotel?

Kode Program 8. Proses *Explatory Data Analysis* pada Python.

```
1. # Filter data reservasi yang dibatalkan
2. canceled_reservations = df[df['is_canceled'] == 1]

3. # Kelompokkan data berdasarkan hotel dan negara, lalu hitung jumlah pembatalan
4. cancellation_by_country = canceled_reservations.groupby(['hotel',
    'country'])['is_canceled'].count().reset_index()

5. # Urutkan data berdasarkan jumlah pembatalan secara menurun
6. cancellation_by_country = cancellation_by_country.sort_values(by=['hotel', 'is_canceled'],
    ascending=[True, False])

7. # Tampilkan negara-negara teratas dengan jumlah pembatalan terbanyak
8. # Tampilkan 10 negara teratas untuk setiap tipe hotel
9. top_cancellations = cancellation_by_country.groupby('hotel').head(5)
```

```

10. # Tampilkan hasilnya
11. display(top_cancellations)

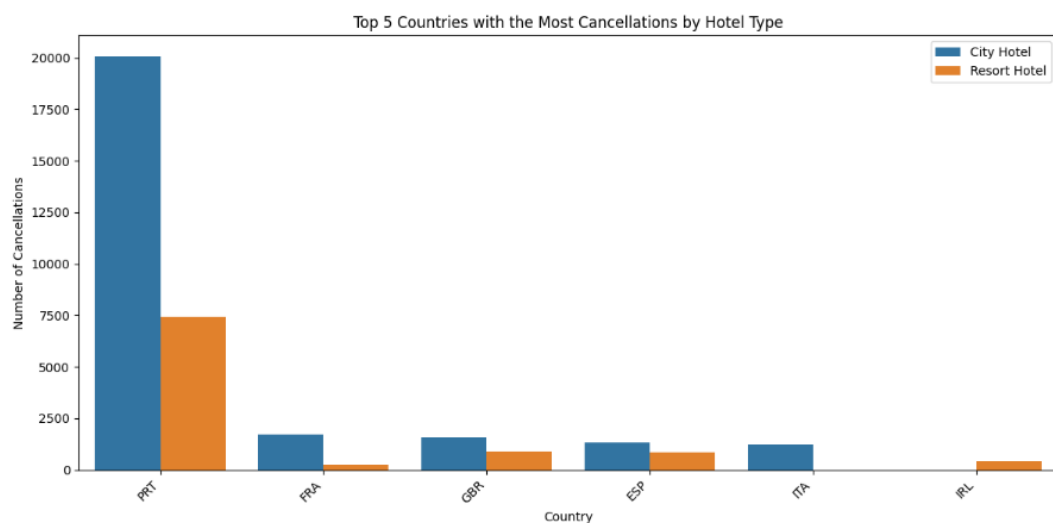
12. # Visualisasi dengan Bar Chart
13. plt.figure(figsize=(12, 6))
14. sns.barplot(x='country', y='is_canceled', hue='hotel', data=top_cancellations)
15. plt.title('Top 5 Countries with the Most Cancellations by Hotel Type')
16. plt.xlabel('Country')
17. plt.ylabel('Number of Cancellations')
18. plt.xticks(rotation=45, ha='right')
19. plt.legend()
20. plt.tight_layout()
21. plt.show()

```

Output:



	hotel	country	is_canceled
96	City Hotel	PRT	20086
37	City Hotel	FRA	1728
40	City Hotel	GBR	1570
33	City Hotel	ESP	1326
59	City Hotel	ITA	1253
182	Resort Hotel	PRT	7448
152	Resort Hotel	GBR	910
147	Resort Hotel	ESP	851
162	Resort Hotel	IRL	432
151	Resort Hotel	FRA	227



Gambar 4.1. Output Kode Program 8.

Wawasan yang di dapatkan:

Berdasarkan dari analisis diatas, dapat disimpulkan bahwa Portugal (PRT) merupakan negara dengan tingkat pembatalan tertinggi, baik untuk City Hotel maupun Resort Hotel. Terdapat perbedaan pola pembatalan antara City Hotel dan Resort Hotel, di mana Portugal mendominasi pembatalan untuk City Hotel, sedangkan untuk Resort Hotel, pembatalan lebih merata dari beberapa negara lain seperti Inggris (GBR), Spanyol (ESP), Irlandia (IRL), dan Prancis (FRA). Fleksibilitas kebijakan pembatalan dan ketidakpastian rencana perjalanan mungkin menjadi faktor-faktor yang berkontribusi terhadap tingginya pembatalan, terutama dari Portugal. Diperlukan investigasi lebih lanjut untuk memahami faktor-faktor spesifik yang menyebabkan tingginya pembatalan dari Portugal dan negara-negara lainnya. Hotel perlu mengembangkan strategi yang spesifik untuk setiap tipe hotel dan negara asal untuk mengurangi pembatalan, seperti penyesuaian kebijakan pembatalan, penawaran insentif, dan komunikasi yang efektif.

IV. Bagaimana perilaku pemesanan Last-Minute yang terjadi?

Kode Program 9. Proses *Explatory Data Analysis* pada Python.

```
1. # Langkah 1: Definisi last-minute booking
2. df['is_last_minute'] = df['lead_time'] <= 7

3. # Langkah 2: Analisis proporsi last-minute booking
4. last_minute_proportion = len(df[df['is_last_minute']]) / len(df) # Perhitungan proporsi
5. print(f"Proporsi pemesanan last-minute: {last_minute_proportion:.2f}") # Menampilkan dengan dua desimal
6. print(f"Proporsi pemesanan last-minute (tanpa pembulatan): {last_minute_proportion}

7. # Visualisasi proporsi last-minute booking dengan pie chart
8. labels = ['Last-Minute Booking', 'Non-Last-Minute Booking']
9. sizes = [last_minute_proportion, 1 - last_minute_proportion] # Proporsi non-last-minute
10. colors = ['#66b3ff', '#ff9999'] # Warna untuk setiap kategori
11. explode = (0.1, 0) # Menonjolkan potongan 'Last-Minute Booking'

12. plt.figure(figsize=(6, 6))
13. plt.pie(sizes, explode=explode, labels=labels, colors=colors, autopct='%1.1f%%', shadow=True,
    startangle=90)
14. plt.title('Proportion of Last-Minute Bookings')
15. plt.axis('equal') # Agar pie chart berbentuk lingkaran
16. plt.show()
```

```
17. # Langkah 3: Analisis tingkat pembatalan
18. # Menghitung tingkat pembatalan untuk last-minute booking dan non-last-minute booking
19. last_minute_cancellation_rate = df[df['is_last_minute']]['is_canceled'].mean() \
20. non_last_minute_cancellation_rate = df[~df['is_last_minute']]['is_canceled'].mean()

21. print(f"Tingkat pembatalan last-minute: {last_minute_cancellation_rate:.2f}")
22. print(f"Tingkat pembatalan non-last-minute: {non_last_minute_cancellation_rate:.2f}")

23. # Visualisasi tingkat pembatalan
24. data = {'Category': ['Last-Minute', 'Non-Last-Minute'],
          'Cancellation Rate': [last_minute_cancellation_rate, non_last_minute_cancellation_rate]}
25. sns.barplot(x='Category', y='Cancellation Rate', data=data)
26. plt.title('Order Cancellation Rate')
27. plt.show()

28. # Langkah 4: Analisis berdasarkan tipe hotel
29. last_minute_by_hotel = df.groupby(['hotel', 'is_last_minute'])['is_canceled'].count().reset_index()
30. sns.barplot(x='hotel', y='is_canceled', hue='is_last_minute', data=last_minute_by_hotel)
31. plt.title('Last-Minute Booking Behavior by Hotel Type')
32. plt.show()

33. # Langkah 5: Analisis berdasarkan bulan
34. # Mengubah tipe data 'arrival_date_month' menjadi kategori dan mengurutkan bulan
35. months_order = ['January', 'February', 'March', 'April', 'May', 'June', 'July', 'August', 'September',
                  'October', 'November', 'December']
36. last_minute_by_month = df.groupby(['arrival_date_month',
                                     'is_last_minute'])['is_canceled'].count().reset_index()
37. last_minute_by_month['arrival_date_month'] = pd.Categorical(last_minute_by_month['arrival_date_month'],
                                                             categories=months_order, ordered=True)
38. last_minute_by_month = last_minute_by_month.sort_values('arrival_date_month')

39. plt.figure(figsize=(12, 6))
40. sns.barplot(x='arrival_date_month', y='is_canceled', hue='is_last_minute', data=last_minute_by_month)
41. plt.title('Last-Minute Booking Behavior by Month')
42. plt.xticks(rotation=45)
43. plt.show()

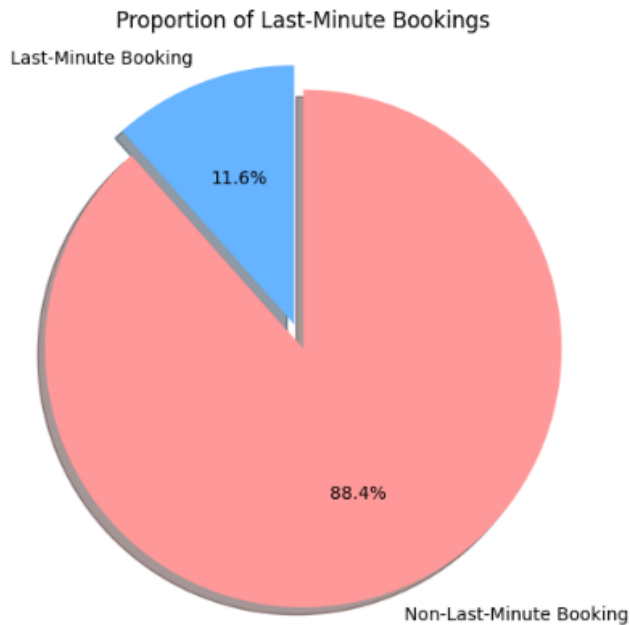
44. # Langkah 6: Analisis berdasarkan negara asal (country)
45. last_minute_by_country = df[df['is_last_minute'] ==
                             True].groupby('country')['is_last_minute'].count().reset_index()
46. last_minute_by_country = last_minute_by_country.sort_values(by=['is_last_minute'], ascending=False)
47. top_10_countries = last_minute_by_country.head(10) # Mengambil 10 negara teratas

48. plt.figure(figsize=(12, 6))
49. sns.barplot(x='country', y='is_last_minute', data=top_10_countries)
50. plt.title('Top 10 Countries with the Most Last-Minute Bookings')
51. plt.xlabel('Country')
```

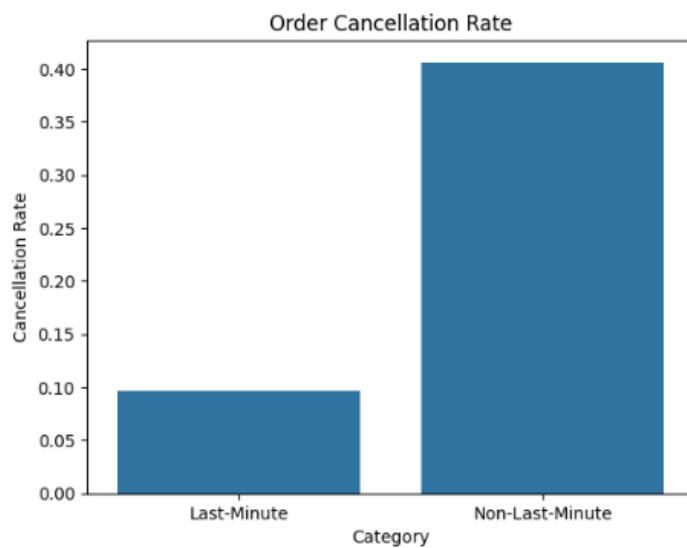
```
52. plt.ylabel('Number of Last-Minute Orders')
53. plt.xticks(rotation=45, ha='right') # Rotasi label sumbu x agar mudah dibaca
54. plt.show()
```

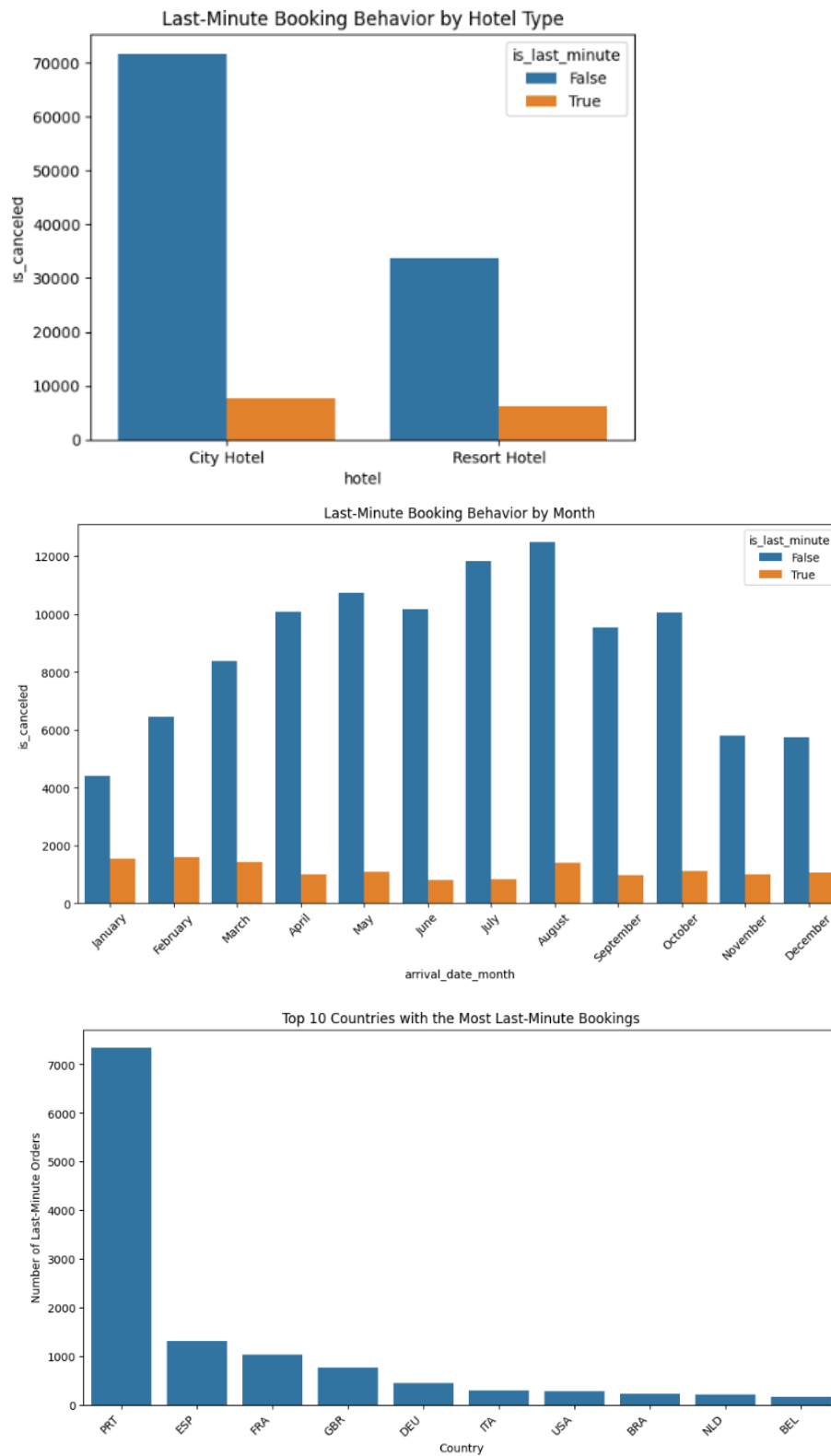
Output:

Proporsi pemesanan last-minute: 0.12
 Proporsi pemesanan last-minute (tanpa pembulatan): 0.1158511048196606



Tingkat pembatalan last-minute: 0.10
 Tingkat pembatalan non-last-minute: 0.41





Gambar 4.1. Output Kode Program 9.

Wawasan yang didapatkan:

Fleksibilitas dan Spontanitas: Pemesanan last-minute mencerminkan fleksibilitas dan spontanitas dalam rencana perjalanan sebagian tamu. Hotel dapat memanfaatkan ini dengan menawarkan penawaran khusus atau promosi untuk pemesanan last-minute.

Pengelolaan Inventaris: Memahami pola pemesanan last-minute dapat membantu hotel dalam mengelola inventaris kamar dengan lebih efisien, terutama di periode dengan tingkat hunian yang rendah.

Strategi Pemasaran: Hotel dapat menyesuaikan strategi pemasaran mereka untuk menargetkan tamu yang cenderung memesan last-minute, misalnya dengan menggunakan online travel agent (OTA) atau media sosial.

Segmentasi Tamu: Dengan menganalisis karakteristik tamu yang melakukan pemesanan last-minute, hotel dapat melakukan segmentasi pasar yang lebih baik dan menawarkan layanan yang sesuai dengan kebutuhan mereka.

Pengurangan Risiko Pembatalan: Meskipun tingkat pembatalan untuk pemesanan last-minute cenderung lebih rendah, hotel tetap perlu menerapkan strategi untuk mengurangi risiko pembatalan, seperti kebijakan pembatalan yang jelas dan komunikasi yang efektif dengan tamu.

Dengan memahami perilaku pemesanan last-minute, hotel dapat mengoptimalkan strategi mereka untuk meningkatkan tingkat hunian, pendapatan, dan kepuasan tamu.

4.2 A/B Testing

Pengujian A/B Testing dalam analisis ini membantu menentukan apakah terdapat perbedaan yang signifikan dalam rata-rata pendapatan harian (ADR) antara City Hotel dan Resort Hotel. Dengan membandingkan kedua kelompok data, pengujian ini memberi bukti apakah salah satu jenis hotel menghasilkan pendapatan lebih tinggi secara signifikan atau tidak. Hasil ini berguna bagi manajemen untuk membuat keputusan berbasis data mengenai penetapan harga, strategi pemasaran, atau kebijakan promosi yang lebih efektif, sehingga dapat meningkatkan profitabilitas dan mengoptimalkan strategi bisnis berdasarkan fakta nyata.

Kode Program 10. Proses pengujian A/B Testing pada Phyton.

```
1. from scipy import stats
2.
3. # Filter data ADR untuk masing-masing jenis hotel
4. city_hotel_adr = df[df['hotel'] == 'City Hotel']['adr']
5. resort_hotel_adr = df[df['hotel'] == 'Resort Hotel']['adr']
6.
7. # Uji t dua sampel independen
8. t_stat, p_value = stats.ttest_ind(city_hotel_adr, resort_hotel_adr)
9.
10. # Cetak hasil uji t dua
11. print(f"Nilai t-statistik: {t_stat:.5f}")
12. print(f"Nilai p-value: {p_value:.5f}")
13.
14. # Tentukan alpha
15. alpha = 0.05
16.
17. # Hasil
18. if p_value < alpha:
19.     print(f"Tolak H0. Ada perbedaan rata-rata ADR antara City Hotel dan Resort
        Hotel (p-value: {p_value:.5f})")
20. else:
    print(f"Gagal menolak H0. Tidak ada perbedaan signifikan dalam rata-rata ADR antara
        City Hotel dan Resort Hotel (p-value: {p_value:.5f})")
```

Penjelasan kode program 5 tentang Proses pengujian A/B Testing pada Phyton, sebagai berikut:

1. Import Library: Menggunakan `scipy.stats` untuk menjalankan uji t dua sampel independen. Library ini menyediakan fungsi yang dapat menghitung nilai t-statistik dan p-value.
2. Filter Data ADR Berdasarkan Jenis Hotel:
 - Baris `city_hotel_adr` mengambil nilai ADR hanya untuk jenis hotel "City Hotel" dari kolom `adr`.
 - Baris `resort_hotel_adr` mengambil nilai ADR hanya untuk jenis hotel "Resort Hotel" dari kolom `adr`.
3. Uji t Dua Sampel Independen:
 - Fungsi `ttest_ind()` digunakan untuk menguji apakah terdapat perbedaan signifikan antara rata-rata ADR dari dua jenis hotel ini.
 - Output berupa nilai t-statistik (`t_stat`) dan p-value (`p_value`) yang digunakan untuk menginterpretasikan hasil.
4. Penentuan Alpha dan Interpretasi Hasil:
 - Nilai alpha ditetapkan sebesar 0,05, atau 5%, yang merupakan batas umum untuk signifikansi statistik.

- Jika p_value kurang dari α , kita menolak hipotesis nol (H_0) dan menyimpulkan bahwa terdapat perbedaan rata-rata ADR yang signifikan antara "City Hotel" dan "Resort Hotel".
- Jika p_value lebih besar dari α , kita gagal menolak H_0 , yang berarti tidak ada perbedaan signifikan dalam rata-rata ADR kedua jenis hotel.

Dari hasil kode program 5 tentang Proses pengujian A/B Testing pada Python akan menghasilkan output yang dapat dilihat pada Gambar 4.2.



```

from scipy import stats

# Filter data ADR untuk masing-masing jenis hotel
city_hotel_adr = df[df['hotel'] == 'City Hotel']['adr']
resort_hotel_adr = df[df['hotel'] == 'Resort Hotel']['adr']

# Uji t dua sampel independen
t_stat, p_value = stats.ttest_ind(city_hotel_adr, resort_hotel_adr)

# Cetak hasil uji t dua
print(f"Nilai t-statistik: {t_stat:.5f}")
print(f"Nilai p-value: {p_value:.5f}")

# Tentukan alpha
alpha = 0.05

# Hasil
if p_value < alpha:
    print(f"Tolak H0. Ada perbedaan rata-rata ADR antara City Hotel dan Resort Hotel (p-value: {p_value:.5f})")
else:
    print(f"Gagal menolak H0. Tidak ada perbedaan signifikan dalam rata-rata ADR antara City Hotel dan Resort Hotel (p-value: {p_value:.5f})")
  
```

Nilai t-statistik: 29.44115
 Nilai p-value: 0.00000
 Tolak H0. Ada perbedaan rata-rata ADR antara City Hotel dan Resort Hotel (p-value: 0.00000)

Gambar 4.2. Output Kode Program 10.

Pada Gambar 4.2 dapat dijelaskan bahwa output dari pengujian A/B Testing menunjukkan bahwa nilai t-statistik sebesar 29.44115 dan p-value mendekati nol (0.00000). Karena p-value jauh lebih kecil dari nilai α yang umumnya digunakan (0.05), kita menolak hipotesis nol (H_0). Ini berarti terdapat perbedaan rata-rata yang signifikan dalam pendapatan harian rata-rata (ADR) antara City Hotel dan Resort Hotel. Hasil ini mengindikasikan bahwa jenis hotel memengaruhi pendapatan harian, dan manajemen dapat mempertimbangkan perbedaan ini dalam strategi penetapan harga atau promosi untuk setiap jenis hotel.

BAB V

Kesimpulan & Saran

A. Kesimpulan utama

Kesimpulan utama dari studi kasus ini adalah bahwa analisis data reservasi pada hotel dapat memberikan kontribusi penting baik dalam ranah ilmu pengetahuan maupun dalam praktik bisnis perhotelan. Melalui studi ini, diperoleh pemahaman mendalam mengenai pola pembatalan, perilaku pemesanan last-minute, perbedaan pola pendapatan antara City Hotel dan Resort Hotel, serta faktor-faktor yang mempengaruhi lead time dan keberhasilan pembayaran. Temuan ini dapat dimanfaatkan oleh manajemen hotel untuk menyusun strategi yang lebih baik dalam meningkatkan kepuasan pelanggan, mengoptimalkan tingkat pendapatan, serta mengelola sumber daya dengan lebih efektif. Dalam praktik bisnis, wawasan ini dapat diterapkan untuk mengembangkan kebijakan yang proaktif, seperti penyesuaian kebijakan pembatalan, peningkatan sistem pembayaran, serta strategi pemasaran yang lebih sesuai dengan karakteristik pelanggan.

B. Saran untuk penelitian selanjutnya

Untuk penelitian selanjutnya, disarankan untuk:

1. **Mengintegrasikan Data Eksternal:** Menambahkan data eksternal, seperti tren wisata global, data ekonomi, dan data cuaca, dapat membantu memahami faktor-faktor eksternal yang memengaruhi tingkat reservasi dan pembatalan. Hal ini akan memperkaya konteks analisis dan memungkinkan pengelola hotel merencanakan strategi berbasis kondisi eksternal.
2. **Analisis Segmentasi Pelanggan:** Melakukan segmentasi yang lebih mendalam berdasarkan demografi, asal negara, serta preferensi pelanggan dapat membantu mengidentifikasi kelompok pelanggan yang paling bernilai. Dengan demikian, hotel dapat mengembangkan layanan yang lebih spesifik sesuai kebutuhan segmen-segmen pelanggan tersebut.

3. **Penerapan Model Prediktif:** Mengembangkan model prediktif berbasis machine learning untuk memprediksi kemungkinan pembatalan reservasi, preferensi waktu pemesanan, serta potensi pendapatan bulanan, memungkinkan hotel mengambil tindakan antisipatif dan personalisasi yang meningkatkan retensi pelanggan.
4. **Pengaruh Faktor Musiman Lebih Mendalam:** Melanjutkan analisis pola musiman secara lebih mendetail dengan mempertimbangkan perbedaan antar bulan atau musim pada berbagai jenis hotel. Hal ini dapat memberikan wawasan tambahan mengenai periode puncak atau sepi, membantu hotel dalam mengelola harga dan ketersediaan kamar.
5. **Eksplorasi Lebih Lanjut tentang Perilaku Last-Minute:** Penelitian mendalam tentang perilaku pelanggan yang sering melakukan pemesanan last-minute, seperti alasan pemesanan mendadak atau faktor yang memengaruhi keputusan pemesanan, bisa membantu dalam menyusun promosi khusus bagi tamu last-minute dan mengoptimalkan strategi penjualan untuk meningkatkan okupansi hotel.

Dengan penelitian lebih lanjut ini, hotel dapat memiliki pemahaman yang lebih komprehensif mengenai faktor-faktor yang memengaruhi operasional dan profitabilitas, serta menerapkan langkah-langkah berbasis data untuk meningkatkan pengalaman pelanggan dan keunggulan kompetitif di pasar.

BAB VI

Lampiran

A. Online Diagram

<https://drive.google.com/file/d/1h-rSXs7ETxVew5O4ViCzskGcdGfyiqzo/view?usp=drivesdk>

B. Python Code

- Business Process Analysis

https://colab.research.google.com/drive/1WsNoCCEco-bckyv7PGxj_hTAgmTM5PUJ?usp=sharing

- Data Pre-processing

<https://colab.research.google.com/drive/1chbHD-2anENHLqvTwge5639whaKHLaXj?usp=sharing>

- Metode A/B Testing

https://colab.research.google.com/drive/1cts3wDDBq-2zUrI4vpgPBgzCgpuDzlYF?usp=chrome_ntp#scrollTo=9QR1tRFac6dm

C. Power Point

<https://docs.google.com/presentation/d/1ZezAUSTTT8dIRszumNEgCtXDPC6jiOq0/edit?usp=sharing&ouid=109353194798525920179&rtpof=true&sd=true>

D. Youtube (Recording Presentasi)

<https://youtu.be/rI9NklZ4R3M>