Data Analytics for Business 2024

# MID EXAM

## Anggota Kelompok


| | | |
|---|---|---|
| 1. | Muhammad Fahmi Hutomo | KM-CS04151 |
| 2. | Marsyanda Nur Zahra | KM-CS04340 |
| 3. | Syahirotul Ambar Maulidiyah | KM-CS04067 |

# BAB I
# Pendahuluan

## 1.1 Latar Belakang Masalah

Perusahaan hotel saat ini berfungsi sebagai penyedia layanan akomodasi yang esensial dalam industri ekonomi dan pariwisata. Namun perusahaan hotel saat ini juga sering menghadapi sejumlah masalah yang dihadapi. Dimana dalam proses reservasi sendiri yang berdampak negatif pada kepuasan pelanggan dan keberlanjutan bisnis dan terdapatnya variasi dalam lead time juga menyebabkan ketidakpastian dalam pengelolaan sumber daya. Dengan memanfaatkan data analitik, hotel dapat memahami pola perilaku pelanggan dan mengidentifikasi masalah yang ada. Dalam rangka mengatasi masalah-masalah tersebut, penting untuk melakukan pemrosesan data secara efektif, menyimpan data dengan cara yang terstruktur, dan melakukan analisis mendalam untuk menemukan insight yang relevan. Tahapan ini mencakup analisis mendalam terhadap pengaruh permintaan khusus, efisiensi proses pembayaran, dan komunikasi selama check-in dan check-out. Hasil dari analisis ini akan digunakan untuk mengembangkan strategi yang dapat meningkatkan kepuasan pelanggan, menurunkan tingkat pembatalan, dan memperbaiki pengalaman secara keseluruhan di hotel. Melalui pendekatan berbasis data ini, perusahaan dapat membangun fondasi yang lebih kuat untuk pertumbuhan jangka panjang.

**Tabel 1.1** Latar Belakang Masalah.

| Modul | Masalah yang Ditemukan |
|---|---|
| Module 1: Business Process Analysis | Proses reservasi yang tidak efisien, tingkat pembatalan yang tinggi, dan kurangnya komunikasi selama check-in dan check-out. |
| Module 2: Python Pre-processing | Data yang tidak konsisten dan perlu dibersihkan, termasuk nilai yang hilang dan format data yang tidak seragam. |
| Module 3: SQL Query | Query yang tidak optimal yang dapat memperlambat pengambilan data, serta kesulitan dalam mengakses informasi yang relevan. |
| Module 4: Python EDA | Kesulitan dalam mengidentifikasi pola dan tren dari data yang kompleks, serta keterbatasan visualisasi untuk analisis mendalam. |
| Module 5: A/B Testing | Tantangan dalam merancang eksperimen yang valid dan memastikan bahwa sampel yang digunakan cukup representatif untuk menarik kesimpulan. |

## 1.2 Tujuan

Tujuan dari penelitian ini adalah untuk menganalisis data pada midterm_hotel_data digunakan untuk mengidentifikasi faktor-faktor yang berkontribusi terhadap masalah yang ada. Melalui analisis yang mendalam, diharapkan dapat ditemukan strategi yang efektif untuk menurunkan tingkat pembatalan dan meningkatkan kepuasan pelanggan. Penelitian ini juga bertujuan untuk memahami pola perilaku pemesanan serta memanfaatkan data dalam pengambilan keputusan yang lebih baik dalam manajemen sumber daya. Selain itu, dengan menerapkan teknik analisis data seperti SQL dan A/B testing, perusahaan dapat mendapatkan wawasan berharga mengenai kinerja hotel dan menguji hipotesis yang dapat mendukung pengembangan strategi bisnis yang lebih baik di masa depan.

## 1.3 Manfaat

Manfaat dari pengerjaan project ini diharapkan perusahaan hotel dapat mengidentifikasi dan mengatasi faktor-faktor yang menyebabkan pembatalan, perusahaan dapat meningkatkan kepuasan pelanggan dan mengurangi kerugian finansial. Melalui pemrosesan dan analisis data, hotel dapat mengembangkan mekanisme yang lebih efektif untuk memenuhi permintaan khusus pelanggan, yang pada akhirnya meningkatkan loyalitas pelanggan. Bagi pelanggan, perbaikan dalam proses reservasi dan layanan akan menghasilkan pengalaman menginap yang lebih memuaskan dan lebih lancar, meningkatkan kemungkinan mereka untuk kembali dan merekomendasikan hotel kepada orang lain. Dengan demikian, penelitian ini tidak hanya memberikan manfaat finansial bagi perusahaan, tetapi juga meningkatkan reputasi dan daya.

# BAB II
# Business Process Analysis

## 2.1 Identifikasi Masalah

Perusahaan hotel mengalami beberapa masalah dalam proses reservasi yang menyebabkan meningkatnya tingkat pembatalan dan rendahnya permintaan pelanggan untuk reservasi jangka panjang. Beberapa masalah yang telah diidentifikasi adalah sebagai berikut:

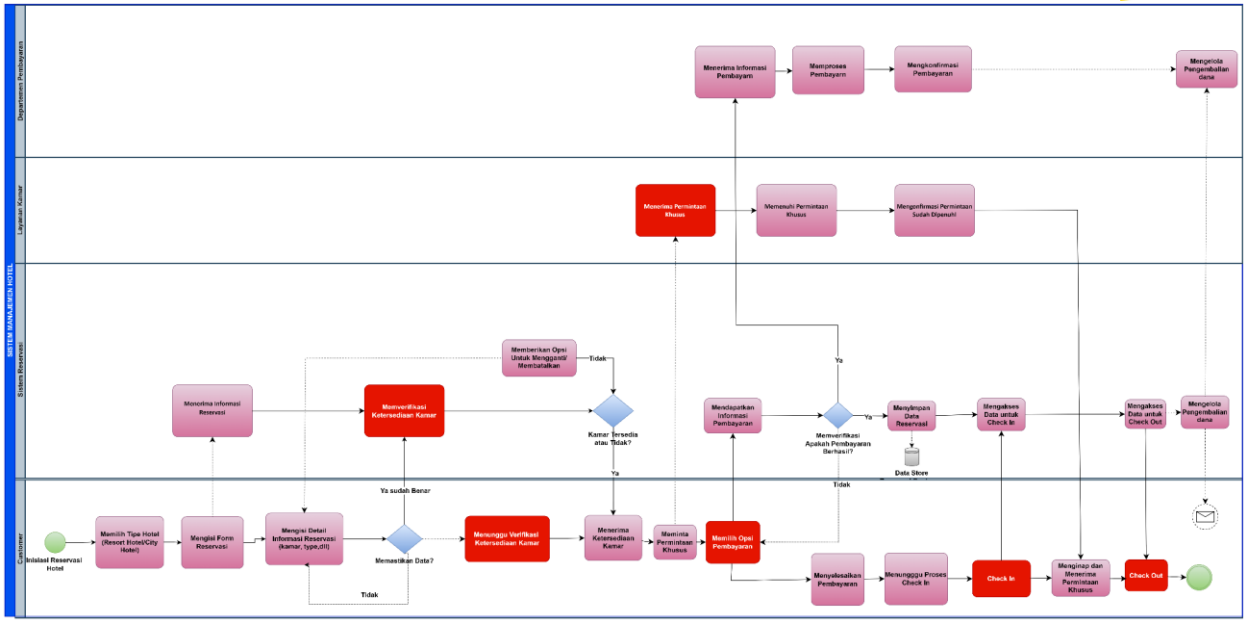
**Tabel 2.1** List Identifikasi Masalah

| Issue (Masalah) | Keterangan | Code (Data Terkait) |
|---|---|---|
| Tingkat Pembatalan yang Tinggi | Persentase pembatalan reservasi cukup tinggi, sering kali pelanggan membatalkan pesanan setelah reservasi. | is_canceled |
| Lead Time yang Bervariasi | Variasi dalam lead time menyebabkan ketidakpastian dalam pengelolaan sumber daya, berpotensi mengakibatkan overbooking atau underbooking. | lead_time |
| Permintaan Khusus Tidak Dikelola dengan Baik | Permintaan khusus dari pelanggan sering kali tidak dipenuhi akibat kurangnya mekanisme penanganan yang efektif, berdampak negatif pada kepuasan pelanggan. | total_of_special_requests |
| Proses Pembayaran yang Tidak Efisien | Kegagalan dalam proses pembayaran membuat pelanggan ragu melanjutkan reservasi, berpotensi meningkatkan pembatalan. | adr, reservation_status |
| Kurangnya Komunikasi Selama Check-in dan Check-out | Informasi tidak diterima tepat waktu oleh pelanggan selama proses check-in dan check-out, menyebabkan ketidakpuasan dan keterlambatan. | reservation_status |

Perusahaan hotel mengalami beberapa masalah dalam proses reservasi yang menyebabkan meningkatnya tingkat pembatalan dan rendahnya permintaan pelanggan untuk reservasi jangka panjang. Beberapa masalah yang telah diidentifikasi adalah sebagai berikut:

## 2.2 Diagram BPMN

Berdasarkan Analisis menunjukkan bahwa perusahaan hotel menghadapi beberapa tantangan dalam proses reservasi yang berdampak signifikan pada kepuasan pelanggan dan kinerja operasional, adapun Tahapan Proses Reservasi dari Awal hingga Akhir beserta masalah yang terkait dijelaskan melalui bpmn berikut.




**Swimlane: Pelanggan**

Proses reservasi hotel dimulai dengan pelanggan yang memilih tipe hotel sesuai preferensi mereka, apakah itu Resort Hotel atau City Hotel. Setelah menentukan pilihan, pelanggan mengisi formulir reservasi dengan data pribadi dan informasi dasar seperti tanggal check-in, check-out, serta jumlah tamu. Pelanggan kemudian memilih detail informasi reservasi, termasuk tipe kamar, jumlah malam, dan paket makanan, sebelum memeriksa kembali data yang telah diisi untuk memastikan semuanya benar. Setelah memastikan keakuratan data, pelanggan menunggu sistem untuk memverifikasi ketersediaan kamar yang mereka pilih. Pelanggan kemudian menerima konfirmasi dari sistem mengenai ketersediaan kamar; jika tidak tersedia, mereka dapat memilih opsi alternatif. Jika diperlukan, pelanggan dapat mengajukan permintaan khusus, memilih metode pembayaran, dan menyelesaikan pembayaran sebelum menunggu tanggal check-in tiba. Pada hari check-in, pelanggan datang ke hotel untuk melakukan proses check-in dan selama masa menginap, mereka menerima layanan sesuai permintaan khusus yang telah diajukan. Setelah selesai menginap, pelanggan melakukan check-out dari hotel dan menyelesaikan administrasi yang diperlukan.

**Swimlane: Sistem Reservasi**

Sistem reservasi bertanggung jawab untuk menerima informasi reservasi yang dimasukkan oleh pelanggan, termasuk tipe hotel, tipe kamar, tanggal, dan informasi tambahan lainnya. Setelah menerima data, sistem memverifikasi ketersediaan kamar untuk tanggal dan

durasi yang diminta oleh pelanggan. Setelah memeriksa ketersediaan, sistem mengirimkan konfirmasi kepada pelanggan. Jika kamar tersedia, proses dilanjutkan ke tahap pembayaran; jika tidak, pelanggan diminta untuk memilih opsi alternatif. Setelah pelanggan memilih metode pembayaran, sistem menerima detail informasi pembayaran dan berkomunikasi dengan Departemen Pembayaran. Sistem kemudian memverifikasi apakah pembayaran berhasil, dan hasil verifikasi ini disampaikan kepada pelanggan. Setelah pembayaran dikonfirmasi berhasil, sistem menyimpan data reservasi untuk keperluan proses check-in dan check-out. Pada hari check-in, sistem memberikan akses kepada petugas hotel untuk mengidentifikasi reservasi pelanggan, dan setelah masa menginap selesai, sistem menyediakan data untuk memproses check-out. Jika ada pembatalan, sistem bekerja sama dengan Departemen Pembayaran untuk mengelola proses pengembalian dana.

**Swimlane: Layanan Kamar**

Layanan Kamar bertugas menerima informasi permintaan khusus dari sistem setelah pelanggan mengajukan permintaan. Setelah menerima permintaan tersebut, layanan kamar memastikan bahwa permintaan khusus, seperti tambahan fasilitas atau layanan tertentu, telah disiapkan sebelum pelanggan check-in atau selama mereka menginap. Setelah memenuhi permintaan tersebut, layanan kamar mengonfirmasi kepada sistem bahwa semua permintaan pelanggan telah dipenuhi dan siap dijalankan.

**Swimlane: Departemen Pembayaran**

Departemen Pembayaran menerima informasi pembayaran dari sistem setelah pelanggan memilih metode pembayaran. Tugas departemen ini adalah memproses transaksi sesuai dengan metode yang dipilih pelanggan. Setelah memproses pembayaran, departemen konfirmasi kepada sistem apakah pembayaran berhasil atau gagal, yang kemudian diteruskan kepada pelanggan. Selain itu, jika ada pembatalan atau situasi yang memerlukan pengembalian dana, Departemen Pembayaran bertanggung jawab untuk memproses pengembalian dana sesuai dengan kebijakan hotel.

Proses reservasi hotel seringkali melibatkan berbagai faktor yang memengaruhi pengalaman pelanggan dan operasional hotel. Dalam analisis ini, kami akan melihat data terkait untuk mengidentifikasi beberapa masalah utama yang dapat mempengaruhi kinerja sistem reservasi. Berikut merupakan analisis masalah berdasarkan data

Analisis Berdasarkan Masalah Berdasarkan BPMN

1. **Tingkat Pembatalan (is_canceled)**
   Tingkat pembatalan dalam proses reservasi ini merupakan metrik penting yang menunjukkan seberapa sering pelanggan membatalkan reservasi setelah melakukan booking.
   Masalah: Jika tingkat pembatalan tinggi, bisa disebabkan oleh ketidakpastian dalam ketersediaan kamar, kebijakan pembatalan yang tidak fleksibel, atau kurangnya kejelasan dalam informasi saat reservasi.
   Solusi: Memperbaiki transparansi dalam ketersediaan kamar dan memberikan kebijakan pembatalan yang lebih fleksibel mungkin dapat mengurangi pembatalan.
2. **Variabilitas Lead Time (lead_time)**
   Lead time mengacu pada jarak waktu antara saat pelanggan membuat reservasi dan tanggal check-in. Variabilitas dalam lead time dapat menjadi indikator masalah ketidakpastian dalam perencanaan pelanggan.
   Masalah: Jika variabilitas lead time terlalu besar, artinya ada ketidakkonsistenan dalam kapan pelanggan melakukan reservasi. Ini dapat memengaruhi perencanaan operasional hotel.
   Solusi: Promosi reservasi awal atau diskon early booking bisa membantu mengurangi variabilitas lead time.
3. **Efektivitas Penanganan Permintaan Khusus (total_of_special_requests)**
   Data permintaan khusus menunjukkan seberapa sering pelanggan mengajukan permintaan tambahan dan bagaimana kemampuan hotel memenuhi permintaan tersebut.
   Masalah: Jika permintaan khusus sering diajukan tetapi tidak dapat dipenuhi secara efektif, ini dapat mengurangi kepuasan pelanggan.
   Solusi: Meningkatkan sistem internal untuk menangani permintaan khusus dengan lebih baik dan memastikan komunikasi yang jelas antara pelanggan dan staf layanan.

**Tabel 2.2** Titik titik Masalah pada Proses BPMN

| Titik Proses | Permasalahan | Dampak yang mungkin terjadi | Solusi yang Diusulkan | Proses dalam BPMN |
|---|---|---|---|---|
| Setelah Konfirmasi Ketersediaan Kamar | Tingkat pembatalan yang tinggi disebabkan oleh ketidakpastian mengenai ketersediaan kamar dan kebijakan pembatalan yang tidak fleksibel. | Mengurangi pendapatan dan meningkatkan biaya operasional. | Meningkatkan transparansi ketersediaan kamar dan menawarkan kebijakan pembatalan yang lebih fleksibel. | "Verifikasi Ketersediaan Kamar" |

| | | | | |
|---|---|---|---|---|
| **Saat Menunggu Verifikasi Ketersediaan Kamar** | Variabilitas dalam lead time menunjukkan ketidakpastian dalam perencanaan, menyebabkan overbooking atau underbooking. | Kesulitan dalam mengelola sumber daya dan pengalaman pelanggan yang tidak konsisten. | Memperkenalkan promosi untuk reservasi awal dan diskon bagi pelanggan yang melakukan pemesanan lebih awal. | "Tunggu Verifikasi" |
| **Setelah Pengajuan Permintaan Khusus** | Permintaan khusus sering tidak dipenuhi karena kurangnya mekanisme penanganan yang efektif. | Menurunkan kepuasan pelanggan dan meningkatkan kemungkinan pembatalan. | Meningkatkan sistem komunikasi internal dan prosedur untuk menangani permintaan khusus secara lebih efektif. | "Terima Permintaan Khusus" |
| **Saat Memilih Metode Pembayaran** | Proses pembayaran yang tidak efisien menyebabkan kegagalan dalam transaksi, membuat pelanggan ragu untuk melanjutkan reservasi. | Mengurangi konversi pemesanan dan meningkatkan frustrasi pelanggan. | Meningkatkan sistem pembayaran untuk memastikan transaksi yang lebih lancar dan memberikan informasi yang jelas. | "Pilih Metode Pembayaran" |
| **Saat Check-in dan Check-out** | Ketidaksempurnaan dalam komunikasi selama check-in dan check-out menyebabkan keterlambatan dan ketidakpuasan pelanggan. | Pengalaman negatif yang dapat merugikan reputasi hotel. | Memperbaiki alur komunikasi antara pelanggan, staf hotel, dan sistem untuk memastikan informasi diterima tepat waktu. | "Proses Check-in" dan "Proses Check-out" |

# BAB III
# Data Preparation and Structured Query Language

## 3.1 Data Preparation

Data preparation dilakukan menggunakan Python dan package Pandas. Proses ini dilakukan di Google Colab. Dataset midterm_hotel_data dimuat dari Google Drive untuk selanjutnya dilakukan proses pembersihan data. Proses pembersihan data dilakukan untuk menangani nilai hilang, nilai tidak konsisten, dan nilai duplikat. Berikut ini adalah langkah-langkah proses pembersihan data yang dilakukan:

1. Import dataset midterm_hotel_data.csv dari google drive
2. Baca dataset dan tampilkan informasi umum dataset (jumlah baris, kolom, dll.).
3. Tampilkan 3 baris pertama dari dataset untuk melihat contoh data.
4. Ubah tipe data kolom `reservation_status_date` menjadi tipe data datetime.
5. Identifikasi data tidak konsisten pada kolom-kolom numerik.
6. Identifikasi data tidak konsisten pada kolom-kolom bertipe objek.
7. Daftar kolom dengan nilai kosong (missing values).
8. Tampilkan info statistik deskriptif untuk kolom-kolom dengan nilai kosong, kecuali kolom `company`.
9. Penanganan missing values untuk kolom `children` dengan mengisi nilai kosong menggunakan `0`.
10. Penanganan missing values untuk kolom `country` dengan mengisi nilai kosong menggunakan nilai modus (nilai yang paling sering muncul).
11. Penanganan missing values untuk kolom `agent` dengan mengisi nilai kosong menggunakan `0`.

12. Identifikasi dan hapus nilai negatif pada kolom `adr` (Average Daily Rate).

13. Isi nilai NaN pada kolom `adr` jika `deposit_type = 'No Deposit'` dengan `0`.

14. Penanganan missing values pada kolom tertentu (`lead_time`, `stays_in_weekend_nights`, `adults`, `adr`, dan `total_of_special_requests`) dengan mengisi nilai kosong menggunakan median dari kolom masing-masing.

15. Hapus duplikat data jika ada, dan reset indeks.

16. Hapus kolom yang tidak diperlukan (`Unnamed: 0`, `company`, `name`, `email`, `phone-number`, dan `credit_card`).

17. Simpan dataset yang sudah bersih sebagai file baru.

Potongan *code* yang digunakan dalam proses data cleansing terlihat seperti di bawah ini.

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

```
# -*- coding: utf-8 -*-
"""midterm-bitlabs-preprocessing.ipynb

Automatically generated by Colab.

Original file is located at
    https://colab.research.google.com/drive/1LF4hPEzwzT78mFPNT9WjuygvxM33ZOun
"""

from google.colab import drive
drive.mount('/content/drive')

import pandas as pd
file_path = '/content/drive/My Drive/MSIB Bitlabs Data Analytics for Business/midterm/midterm_hotel_data.csv'
dataset = pd.read_csv(file_path)

dataset.info()

dataset.head(3)

# mengubah tipe data reservation_status_date menjadi datetime untuk konsistensi
dataset['reservation_status_date'] = pd.to_datetime(dataset['reservation_status_date'])

# mengidentifikasi data tidak konsisten
for col in dataset.describe().columns:
    print(f"Unique values in column '{col}':")
    print(dataset[col].unique())
    print('-'*70)

# mengidentifikasi data tidak konsisten pada kolom-kolom bertipe object
for col in dataset.describe(include='object').columns:
    print(f"Unique values in column '{col}':")
    print(dataset[col].unique())
    print('-'*70)
```

```
35.
36. # mendaftar semua kolom yang mengandung missing values
37. cols_missing_values = dataset.columns[dataset.isna().any()]
38. print(cols_missing_values)
39.
40. dataset[cols_missing_values].info()
41.
42. # membaca statistik deskriptif kolom-kolom yang mengandung missing values untuk memperoleh informasi terkait penanganan missing values kecuali 'company'
43. dataset[['lead_time', 'stays_in_weekend_nights', 'adults', 'children', 'agent', 'adr', 'total_of_special_requests']].describe()
44.
45. dataset['country'].describe()
46.
47. # penanganan missing values pada kolom 'children'
48. dataset.fillna({'children': 0}, inplace=True)
49.
50. # penanganan missing values pada kolom 'country'
51. dataset.fillna({'country': dataset['country'].mode()[0]}, inplace=True)
52.
53. # penanganan missing values pada kolom 'agent'
54. dataset.fillna({'agent': 0}, inplace=True)
55.
56. # penanganan data tidak konsisten pada kolom 'adr' (nilai negatif)
57. negative_adr_indices = dataset[dataset['adr'] < 0].index
58. dataset = dataset.drop(negative_adr_indices).reset_index(drop=True)
59.
60. # mengisi kolom 'adr' == NaN dan deposit_type nya No Deposit dengan 0
61. dataset.loc[dataset['deposit_type'] == 'No Deposit', 'adr'] = dataset.loc[dataset['deposit_type'] == 'No Deposit', 'adr'].fillna(0)
62.
63. dataset[cols_missing_values].info()
64.
65. dataset[['lead_time', 'stays_in_weekend_nights', 'adults', 'adr', 'total_of_special_requests']].describe()
66.
67. # penanganan missing values pada kolom 'lead_time', 'stays_in_weekend_nights', 'adults', 'adr', dan 'total_of_special_requests'
68. columns = ['lead_time', 'stays_in_weekend_nights', 'adults', 'adr', 'total_of_special_requests']
69. for column in columns:
70.     dataset.fillna({column: dataset[column].median()}, inplace=True)
71.
72. dataset.info()
73.
74. # menghapus records duplikat jika ada dan mengatur indeks baru
75. dataset.drop_duplicates(inplace=True)
76. dataset.reset_index(drop=True, inplace=True)
77. dataset.info()
78.
79. # menghapus kolom-kolom yang tidak diperlukan untuk analisis
80. dataset = dataset.drop(columns=['Unnamed: 0', 'company', 'name', 'email','phone-number', 'credit_card'])
81.
82. dataset.head()
83.
84. dataset.info()
85.
86. dataset.to_csv('data_hotel_clean.csv', index=False)
```

## 3.2 Data Extraction

Pada tahap data extraction ini dilakukan beberapa queri untuk memperoleh beberapa insight di antaranya:

- Pembatalan pembatalan lebih banyak dilakukan di City Hotel dari tahun ke tahun.

**Kode Program 2**. Persentase Pembatalan Reservasi untuk setiap hotel per Tahun

```
1.  SELECT
2.     hotel,
3.     arrival_date_year,
4.     (SUM(is_canceled) * 100.0 / COUNT(*)) AS cancellation_rate
5.  FROM
6.     data_hotel_clean
7.  GROUP BY
8.     arrival_date_year, hotel
9.  ORDER BY
10.    hotel, arrival_date_year;
```

Berikut hasil ouput dari presentase pembatalan untuk setiap hotel per tahun

| hotel | arrival_date_year | cancellation_rate |
|---|---|---|
| City Hotel | 2,015 | 43.8824733226 |
| Resort Hotel | 2,015 | 25.715660332 |
| City Hotel | 2,016 | 40.395909806 |
| Resort Hotel | 2,016 | 26.5524855927 |
| City Hotel | 2,017 | 42.5003635306 |
| Resort Hotel | 2,017 | 30.7633935347 |

- Pemesanan dengan rata-rata lead_time paling singkat memiliki permintaan khusus terbanyak.

**Kode Program 3**. Rata-rata lead_time untuk lead_time lebih dari rata-rata lead_time keseluruhan untuk setiap Total Pemesanan Khusus

```
11. WITH overall_avg_lead_time AS (
12.    SELECT AVG(lead_time) AS avg_lead_time
13.    FROM data_hotel_clean
14. )
15. SELECT
16.    total_of_special_requests,
17.    AVG(lead_time) AS average_lead_time
18. FROM
19.    data_hotel_clean
20. WHERE
21.    lead_time > (SELECT avg_lead_time FROM overall_avg_lead_time)
22. GROUP BY
23.    total_of_special_requests
24. ORDER BY
25.    total_of_special_requests;
```

Berikut hasil ouput dari Rata-rata lead_time untuk lead_time lebih dari rata-rata lead_time keseluruhan untuk setiap Total Pemesanan Khusus

| | total_of_special_requests | average_lead_time |
|---|---|---|
| 1 | 0 | 210.7666372259 |
| 2 | 1 | 190.4127382146 |
| 3 | 2 | 182.4735099338 |
| 4 | 3 | 181.7979381443 |
| 5 | 4 | 197.0416666667 |
| 6 | 5 | 171.5454545455 |

- Pendapatan tertinggi untuk pemesanan hotel dengan permintaan khusus lebih dari 2 terjadi pada tanggal 17 Agustus 2017 dengan total adr 2.656,86.

**Kode Program 4**. Tanggal ketika Diperoleh Pendapatan Tertinggi

```
26. SELECT
27.     reservation_status_date,
28.     SUM(adr) AS total_revenue
29. FROM
30.     data_hotel_clean
31. WHERE
32.     reservation_status = 'Check-Out'
33.     AND total_of_special_requests > 2
34. GROUP BY
35.     reservation_status_date
36. ORDER BY
37.     total_revenue DESC
38. LIMIT 1;
39.
```

Berikut hasil ouput dari Tanggal ketika Diperoleh Pendapatan Tertinggi

| reservation_status_date | total_revenue |
|---|---|
| 2017-08-17 | 2,656.86 |

- Pemesanan dengan keberhasilan pembayaran tertinggi berada di rentang lead_time antara 70 sampi 161 hari (di antara quartil kedua dan kuartil ketiga)

**Kode Program 5**. Hubungan lead_time di Rentang Tertentu Dengan Keberhasilan Pembayaran

```
40. WITH lead_time_groups AS (
41.     SELECT
42.         CASE
43.             WHEN lead_time BETWEEN 0 AND 18 THEN '0-18 days' /* 18 = q1*/
44.             WHEN lead_time BETWEEN 19 AND 69 THEN '19-69 days' /*69 = q2*/
45.             WHEN lead_time BETWEEN 70 AND 161 THEN '70-161 days' /*161 = q3*/
46.             WHEN lead_time > 161 THEN '>161 days'
47.         END AS lead_time_range,
48.         adr
49.     FROM
```

```
        data_hotel_clean
)
SELECT
    lead_time_range,
    AVG(adr) AS average_adr
FROM
    lead_time_groups
GROUP BY
    lead_time_range
ORDER BY
    lead_time_range;
```

Hubungan lead_time di Rentang Tertentu Dengan Keberhasilan Pembayaran

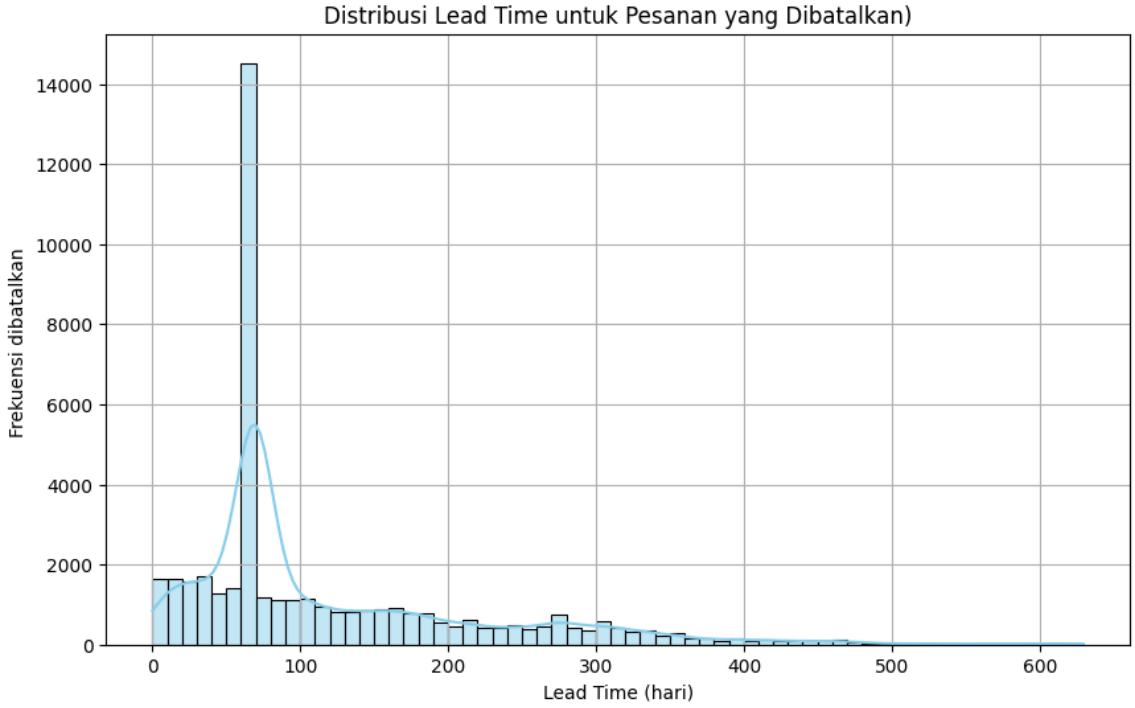
| lead_time_range | average_adr |
|---|---|
| 0-18 days | 66.8706131408 |
| 19-69 days | 75.4131736569 |
| 70-161 days | 78.4843855871 |
| >161 days | 74.053963127 |

# BAB IV
# Python Programming

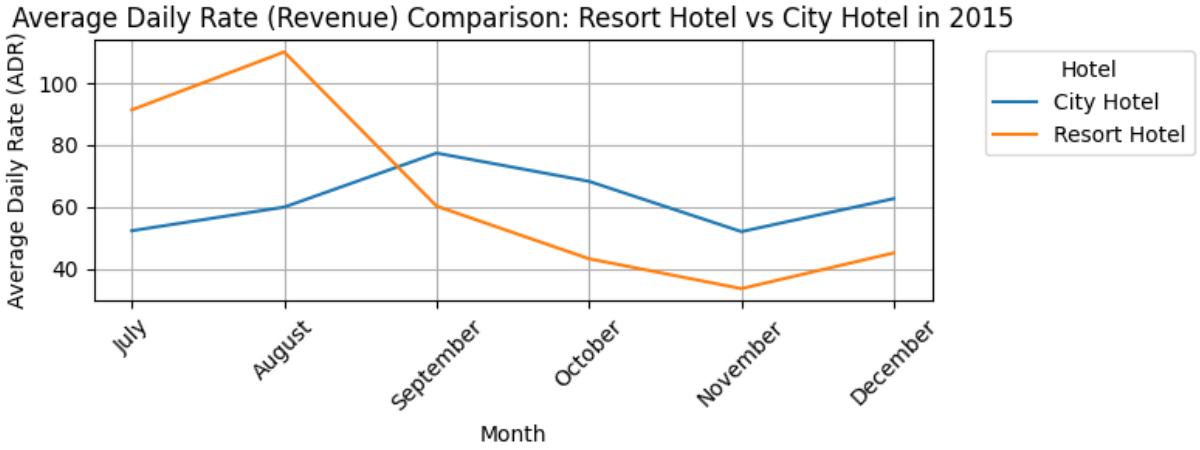
## 4.1 Exploratory Data Analysis

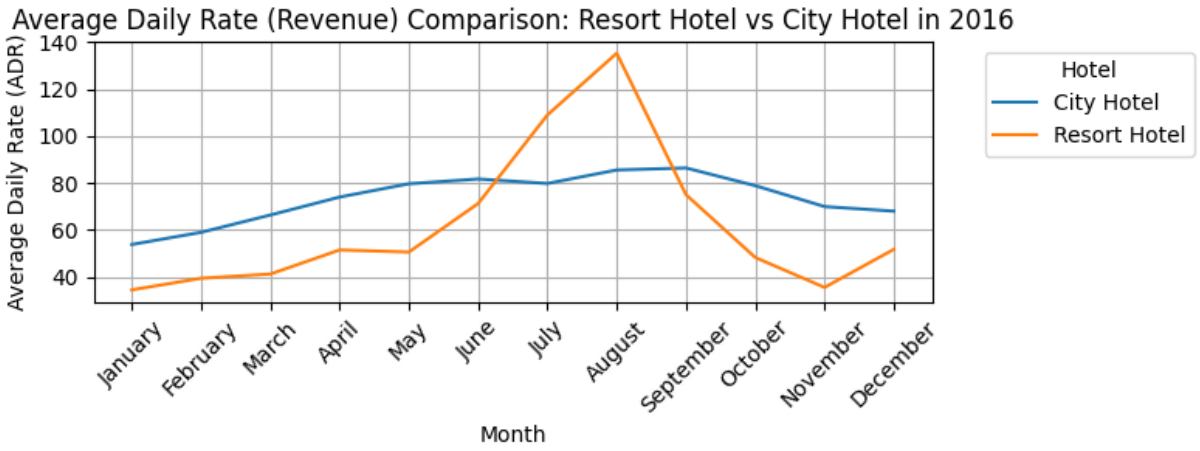
Pada tahap ini diperoleh beberapa insight di antaranya:

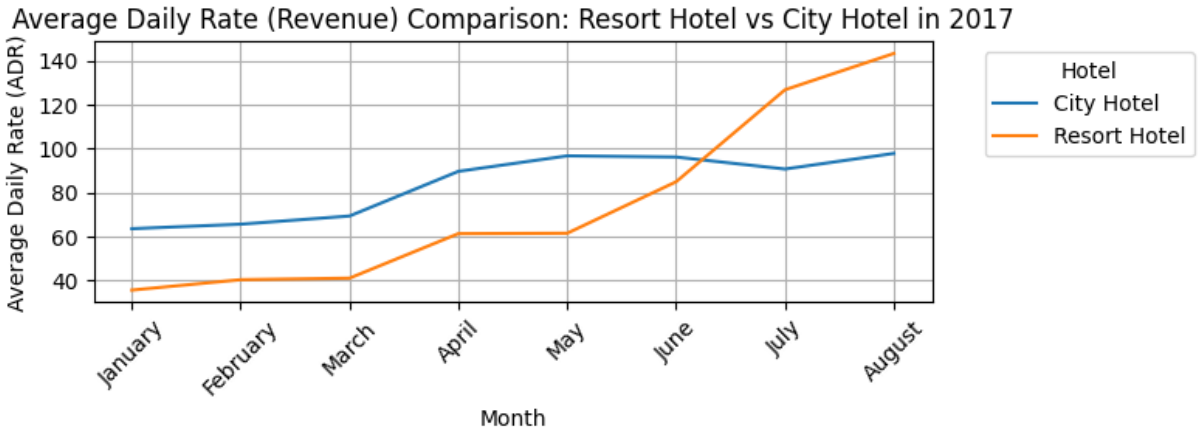
- Pembatalan pesanan cenderung lebih banyak terjadi untuk pemesanan dengan lead_time pendek. Kecuali untuk lead_time dengan jangka waktu di antara 60 hingga 70 hari. Pada rentang ini terjadi banyak sekali pembatalan jika dibandingkan dengan rentang hari yang lain.



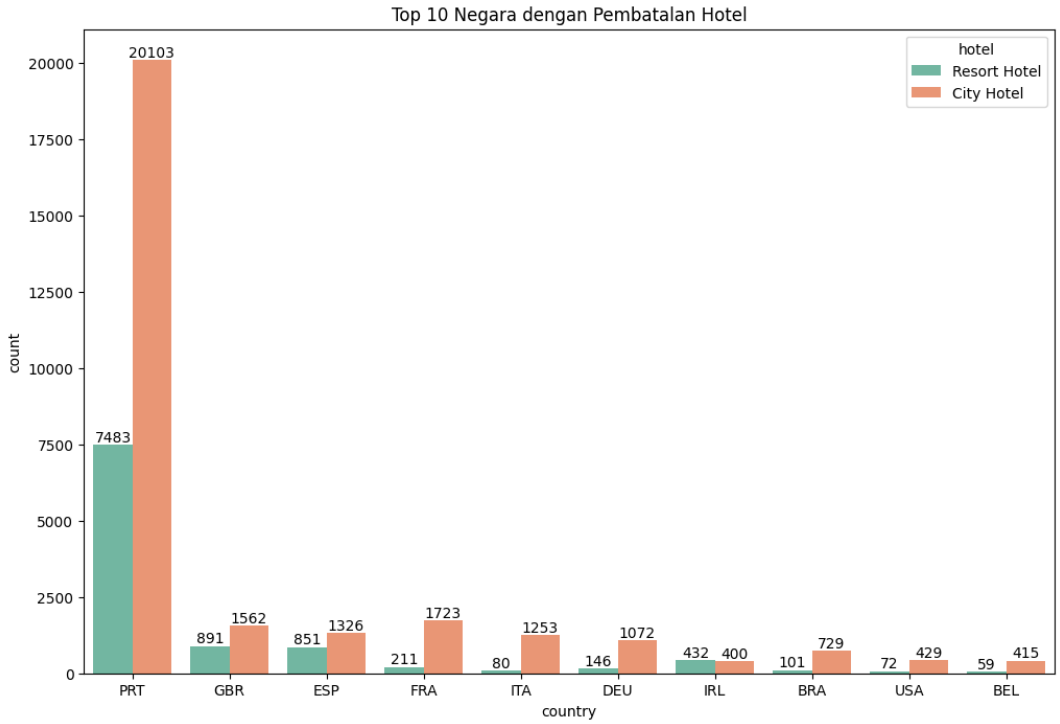

- Dari awal Juni hingga akhir Agustus pendapatan Resort Hotel lebih tinggi dibanding City Hotel. Namun pendapatan City Hotel lebih besar dibanding Resort Hotel dari akhir Agustus hingga awal Juni.



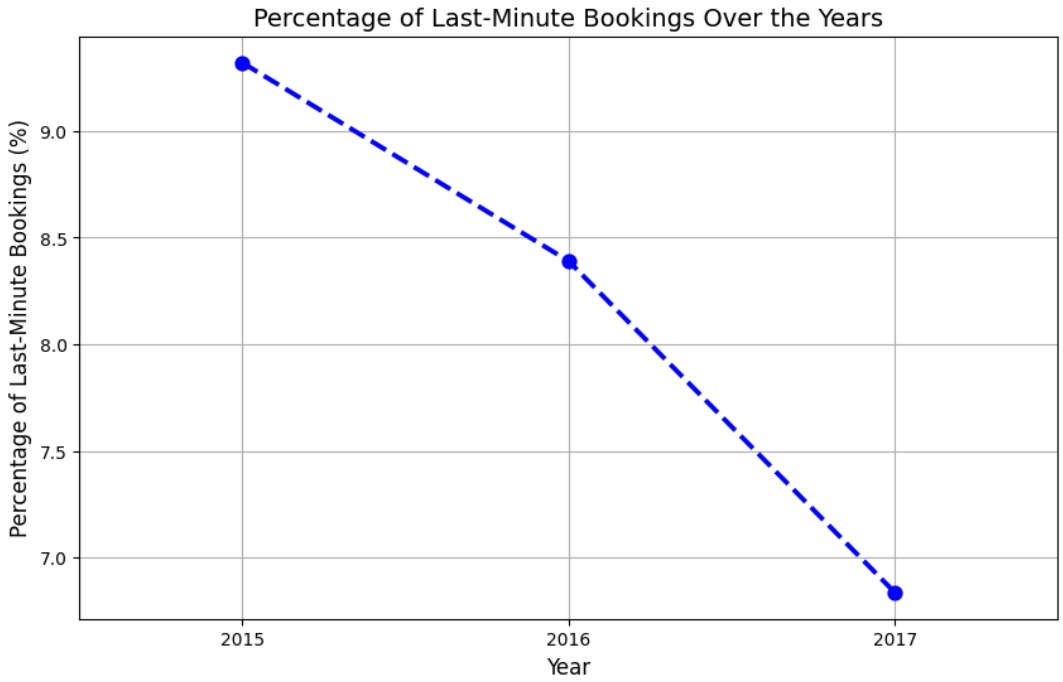




- Turis dari negara berinisial PRT paling banyak melakukan pembatalan baik di Resort Hotel maupun di City Hotel




- Pemesanan last-minute memiliki tingkat pembatalan yang kecil selama dua tahun terakhir ( 8,47%).

- Persentase pemesanan las-minute mengalami penurunan dari tahun ke tahun.




Source code untuk grafik-grafik di atas dapat dilihat di bawah ini.

**Kode Program 5.** Hubungan Lead_time dengan Jumlah Pesanan yang Dibatalkan

```
# -*- coding: utf-8 -*-
"""modul4_EDA.ipynb

Automatically generated by Colab.

Original file is located at
    https://colab.research.google.com/drive/1VmvCC5_kLIIQGJaaaSq-k3sN9fQYuDt-
"""

from google.colab import drive
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

drive.mount('/content/drive')

file_path = '/content/drive/My Drive/MSIB Bitlabs Data Analytics for Business/midterm/data_hotel_clean.csv'
data = pd.read_csv(file_path)

for col in data.describe().columns:
    print(f"Unique values in column '{col}':")
    print(data[col].unique())
    print('-'*70)

for col in data.describe(include='object').columns:
    print(f"Unique values in column '{col}':")
    print(data[col].unique())
    print('-'*70)

"""
```

```
Modul 4 nomor 1

1. Bagaimana rasio antara lead time dengan jumlah hari reservasi dibatalkan (is_canceled == 1)
"""

canceled_data = data[data['is_canceled'] == 1]
plt.figure(figsize=(10, 6))
sns.histplot(canceled_data['lead_time'], bins=63, kde=True, color='skyblue')
plt.title('Distribusi Lead Time untuk Pesanan yang Dibatalkan)')
plt.xlabel('Lead Time (hari)')
plt.ylabel('Frekuensi dibatalkan ')
plt.grid(True)
plt.show()



"""## Modul 4 nomor 2

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

monthly_revenue = data.groupby(['hotel', 'arrival_date_year', 'arrival_date_month'])['adr'].mean().reset_index()

# Convert the month name into a categorical type to ensure proper order in the plot
monthly_revenue['arrival_date_month'] = pd.Categorical(
    monthly_revenue['arrival_date_month'],
    categories=['January', 'February', 'March', 'April', 'May', 'June', 'July', 'August', 'September', 'October', 'November', 'December'],
    ordered=True
)

# Creating separate plots for each year
years = [2015, 2016, 2017]

plt.figure(figsize=(8, 9))

for i, year in enumerate(years, 1):
    plt.subplot(3, 1, i)
    sns.lineplot(
        data=monthly_revenue[monthly_revenue['arrival_date_year'] == year],
        x='arrival_date_month',
        y='adr',
        hue='hotel',
        markers=True,
        dashes=False
    )
    plt.title(f'Average Daily Rate (Revenue) Comparison: Resort Hotel vs City Hotel in {year}')
    plt.xlabel('Month')
    plt.ylabel('Average Daily Rate (ADR)')
    plt.grid(True)
    plt.xticks(rotation=45)
    plt.legend(title='Hotel', bbox_to_anchor=(1.05, 1), loc='upper left')

plt.tight_layout()
plt.show()

"""## Modul 4 nomor 3

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

dataset = data[data['is_canceled'] == 1]
top_10_countries = dataset['country'].value_counts().nlargest(10).index
```

```
filtered_dataset = dataset[dataset['country'].isin(top_10_countries)]

plt.figure(figsize=(12, 8))
ax = sns.countplot(x='country', data=filtered_dataset, hue='hotel', palette='Set2', order = top_10_countries)

for container in ax.containers:
            ax.bar_label(container)

        plt.title('Top 10 Negara dengan Pembatalan Hotel')
        plt.show()

        city_dataset = dataset[dataset['hotel'] == 'City Hotel']
        top_10 = city_dataset['country'].value_counts().nlargest(10).index
        filtered_city_dataset = city_dataset[city_dataset['country'].isin(top_10)]

        plt.figure(figsize=(8, 6))
        ax_city     =     sns.countplot(x='country',     data=filtered_city_dataset, palette='Set2', order=top_10)
        ax_city.set_title('Top 10 Negara dengan Pembatalan di City Hotel')

        for container in ax_city.containers:
            ax_city.bar_label(container)

        plt.show()

        resort_dataset = dataset[dataset['hotel'] == 'Resort Hotel']
        top_10 = resort_dataset['country'].value_counts().nlargest(10).index
        filtered_resort_dataset                                                 = resort_dataset[resort_dataset['country'].isin(top_10)]

        plt.figure(figsize=(8, 6))
        ax = sns.countplot(x='country', data=filtered_resort_dataset, palette='Set2', order=top_10)
        ax.set_title('Top 10 Negara dengan Pembatalan di Resort Hotel')
        for container in ax.containers:
            ax.bar_label(container)

        plt.show()

        """ ##Modul 4 Nomor 4
        4. Bagaimana perilaku pemesanan Last-Minute yang terjadi?
        """

        last_minute_threshold = 3
        data['is_last_minute'] = data['lead_time'] <= last_minute_threshold

        canceled_last_minute                                                    = data[data['is_last_minute']].groupby('is_canceled').size()                            / data[data['is_last_minute']].shape[0] * 100

        plt.figure(figsize=(8, 6))
        ax        =        sns.barplot(x=['Not        Canceled',        'Canceled'], y=canceled_last_minute.values)
        for container in ax.containers:
            ax.bar_label(container,  fmt='%.2f%%')

        plt.title('Canceled Rate for Last-Minute Bookings', fontsize=14)
        plt.xlabel('Reservation Status', fontsize=12)
        plt.ylabel('Percentage (%)', fontsize=12)
        plt.grid(True)
        plt.show()


        last_minute_threshold = 3 # last minute =
        data['is_last_minute'] = data['lead_time'] <= last_minute_threshold
        last_minute_by_year                                                     = data.groupby('arrival_date_year')['is_last_minute'].mean() * 100
```

```
154.
155.        plt.figure(figsize=(10, 6))
156.        sns.pointplot(x=last_minute_by_year.index,      y=last_minute_by_year.values,
    markers="o", linestyles="--", color="blue")
157.        plt.title('Percentage of Last-Minute Bookings Over the Years', fontsize=14)
158.        plt.xlabel('Year', fontsize=12)
159.        plt.ylabel('Percentage of Last-Minute Bookings (%)', fontsize=12)
160.        plt.grid(True)
161.        plt.show()
```

## 4.2 **A/B Testing**

Uji statistik T-Test dilakukan untuk mengevaluasi hipotesis bahwa terdapat perbedaan signifikan dalam rata-rata Average Daily Rate (ADR) antara City Hotel dan Resort Hotel. Variabel yang digunakan dalam pengujian ini adalah kolom 'hotel' dan kolom 'adr'.

Hipotesis yang diuji dalam analisis ini adalah sebagai berikut:

- Hipotesis Nol (H0): Tidak ada perbedaan signifikan dalam rata-rata ADR antara City Hotel dan Resort Hotel.

- Hipotesis Alternatif (H1): Terdapat perbedaan signifikan dalam rata-rata ADR antara City Hotel dan Resort Hotel.

Berdasarkan hasil T-Test, diperoleh nilai statistik T dan p-value sebagai berikut:

T-statistic: 26.303561136359946

P-value: 4.752950071579652e-152

Karena p-value lebih kecil dari tingkat signifikansi 0.05, maka kami menolak hipotesis nol. Hasil ini mendukung adanya perbedaan yang signifikan antara rata-rata Average Daily Rate dari City Hotel dan Resort Hotel. Perbedaan ini menunjukkan bahwa faktor jenis hotel mungkin mempengaruhi rata-rata ADR.

**Kode Program 5**. A/B Testing menggunakan T-test

```
# -*- coding: utf-8 -*-
"""t-test.ipynb

Automatically generated by Colab.

Original file is located at
```

```
https://colab.research.google.com/drive/1NApvOfiuXjVbqpEw4mbeZ4moeiXIQufC
"""

from google.colab import drive
drive.mount('/content/drive')

import pandas as pd
file_path = '/content/drive/My Drive/MSIB Bitlabs Data Analytics for Business/midterm/data_hotel_clean.csv'
data = pd.read_csv(file_path)

print('H0 = Tidak ada perbedaan signifikan dalam rata-rata Average Daily Rate (ADR) antara City Hotel dan Resort Hotel \nH1 = Terdapat perbedaan signifikan dalam rata-rata Average Daily Rate (ADR) antara City Hotel dan Resort Hotel')

from scipy import stats

city_hotel_adr = data[data['hotel'] == 'City Hotel']['adr']
resort_hotel_adr = data[data['hotel'] == 'Resort Hotel']['adr']

# Melakukan t-test
t_stat, p_value = stats.ttest_ind(city_hotel_adr, resort_hotel_adr)

# Menampilkan hasil
print(f"T-statistic: {t_stat}")
print(f"P-value: {p_value}")

if p_value < 0.05:
    print("Tolak H0: Terdapat perbedaan signifikan dalam ADR antara City Hotel dan Resort Hotel.")
else:
    print("Gagal menolak H0: Tidak terdapat perbedaan signifikan dalam ADR antara City Hotel dan Resort Hotel.")
```

# BAB V
# Kesimpulan & Saran

## A. Kesimpulan

Tingkat pembatalan pemesanan di Resort Hotel dan City Hotel cukup tinggi, yakni lebih dari 25% untuk Resort Hotel dan lebih dari 40% untuk City Hotel setiap tahunnya. Namun hal yang cukup jauh berbeda terjadi pada pemesanan last-minute yang mana persentase pembatalan pesanannya hanya mencapai 8,47%.

Jumlah pembatalan pesanan secara umum lebih banyak terjadi pada lead_time yang pendek, kecuali pada pemesanan dengan lead_time di antara 60-70 hari yang mana pada rentang ini terjadi sangat banyak pembatalan dibandingkan dengan rentang yang lain.

Turis-turis yang banyak melakukan pembatalan kebanyakan berasal dari benua Eropa dan negara dengan inisial PRT menjadi negara dengan turis yang melakukan pembatalan terbanyak baik di Resort Hotel maupun City Hotel.

Setelah dilakukan segmentasi lead_time berdasarkan kuartil, diketahui bahwa rata-rata ADR tertinggi berada pada rentang lead_time di antara q1 dan q2, yaitu sebesar 78,84. Angka ini jauh di bawah ADR tertinggi yang didapatkan pada tanggal 17 Agustus 2017 sebesar 2.656,86.

Pendapatan Resort Hotel dan City Hotel memiliki pola yang sama pada bulan-bulan tertentu di mana pada awal Juni hingga akhir Agustus pendapatan Resort Hotel selalu lebih besar dari City Hotel , tetapi hal sebaliknya terjadi pada bulan-bulan yang lain, sehingga dapat dikatakan bahwa Resort Hotel memperoleh pendapatan lebih tinggi dibanding City Hotel selama musim panas.

## B. Saran

Penelitian selanjutnya perlu dilakukan terkait customer_type, deposit_type, market_segment, dan distribution_channel untuk lebih memahami pola khas customer yang melakukan pembatalan dan untuk mengetahui sebab-sebab yang mungkin membuat mereka membatalkan pesanan.

City Hotel perlu mempertimbangkan fitur-fitur pelayanan terkait musim panas yang ada pada Resort Hotel untuk meningkatkan pendapatan selama musim panas.

# BAB VI
# Lampiran

## A. Online Diagram

BPMN
https://drive.google.com/file/d/1eorZnUw-xT-4jTECaNPzBULCnLU4rgQ7/view?usp=sharing

## B. Python Code

Data Preparation:
https://colab.research.google.com/drive/1LF4hPEzwzT78mFPNT9WjuygvxM33ZOun?usp=sharing
EDA:
https://colab.research.google.com/drive/1VmvCC5_kLIIQGJaaaSq-k3sN9fQYuDt-?usp=sharing
T-Test:
https://colab.research.google.com/drive/1NApvOfiuXjVbqpEw4mbeZ4moeiXIQufC?usp=sharing

## C. Recording

1. Link Recording
   https://youtu.be/4J9uP3cBJGA?si=y16aHv5zJE-ajwHZ