



Muhammad Shobri Al Mughdhor

DIGITAL SKILL FAIR 39.0: DATA - DATA SCIENCE

Hotel Booking Demand Datasets

Muhammad Shobri Al Mughdhor



Shobri24

Dataset

Dataset ini berisi informasi pemesanan dari dua jenis hotel—city hotel dan resort hotel—with total 51.390 entri dan 32 kolom yang mencakup variabel numerik (seperti lead time, ADR) dan kategorik (seperti meal type, market segment). Data asli berasal dari Hotel Booking Demand Datasets (Antonio et al., 2019) yang telah dibersihkan oleh Mock & Bichat (2020), sehingga siap untuk dianalisis lebih lanjut. Tujuan utamanya adalah melakukan Handling Missing Value dan Duplicate.



Exploratory Data

Missing Value

Dataset ini mengandung sejumlah missing value (nilai kosong) yang tersebar di beberapa kolom. Pada kolom children, terdapat 4 data yang hilang, sementara kolom country memiliki 478 missing value. Missing value yang lebih signifikan muncul pada kolom agent dengan 8.864 data kosong dan kolom company yang mencapai 48.031 missing value. Selain itu, beberapa kolom lain seperti required_car_parking_spaces, total_of_special_requests, reservation_status, dan reservation_status_date masing-masing memiliki 1 missing value.

children

4

babies

0

meal

0

country

478

agent

8864

company

48031

days_in_waiting_list

0

customer_type

0

adr

0

required_car_parking_spaces

1

total_of_special_requests

1

reservation_status

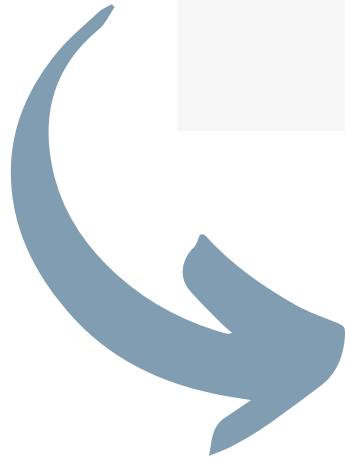
1

reservation_status_date

1

Exploratory Data

```
for kolom in data.columns:  
    if kolom in ['country', 'reservation_status_date', 'reservation_status']:  
        # Untuk kolom bertipe object teks/kategori, isi dengan nilai modus (nilai paling sering muncul)  
        data[kolom] = data[kolom].fillna(data[kolom].mode()[0])
```



Mengatasi Missing Value

Untuk memastikan kelengkapan data, dilakukan penanganan missing value dengan pendekatan yang berbeda berdasarkan jenis kolomnya. Pada kolom bertipe teks/kategori seperti country, reservation_status_date, dan reservation_status, nilai kosong diisi dengan modus (nilai paling sering muncul) karena nilai ini paling mewakili distribusi data kategorikal. Sementara itu, untuk kolom numerik seperti children, agent, company, required_car_parking_spaces, dan total_of_special_requests, nilai kosong diisi dengan nilai rata-rata agar tidak mengganggu analisis statistik.

```
elif kolom in ['children', 'agent', 'company', 'required_car_parking_spaces', 'total_of_special_requests']:  
    # Untuk kolom float numerik, isi dengan nilai rata-rata  
    data[kolom] = data[kolom].fillna(data[kolom].mean())
```



Exploratory Data

Mengecek Duplicate

Setelah dilakukan pengecekan duplikasi data, ditemukan sebanyak 11.050 entri yang merupakan data ganda (duplikat) dari total 51.390 entri dalam dataset. Angka ini menunjukkan bahwa sekitar 21,5% dari seluruh data merupakan duplikat, suatu jumlah yang signifikan dan perlu ditangani secara serius.

```
check_duplicate = data.duplicated().sum()

print(f"Jumlah data yang duplikat = {check_duplicate}")

Jumlah data yang duplikat = 11050
```

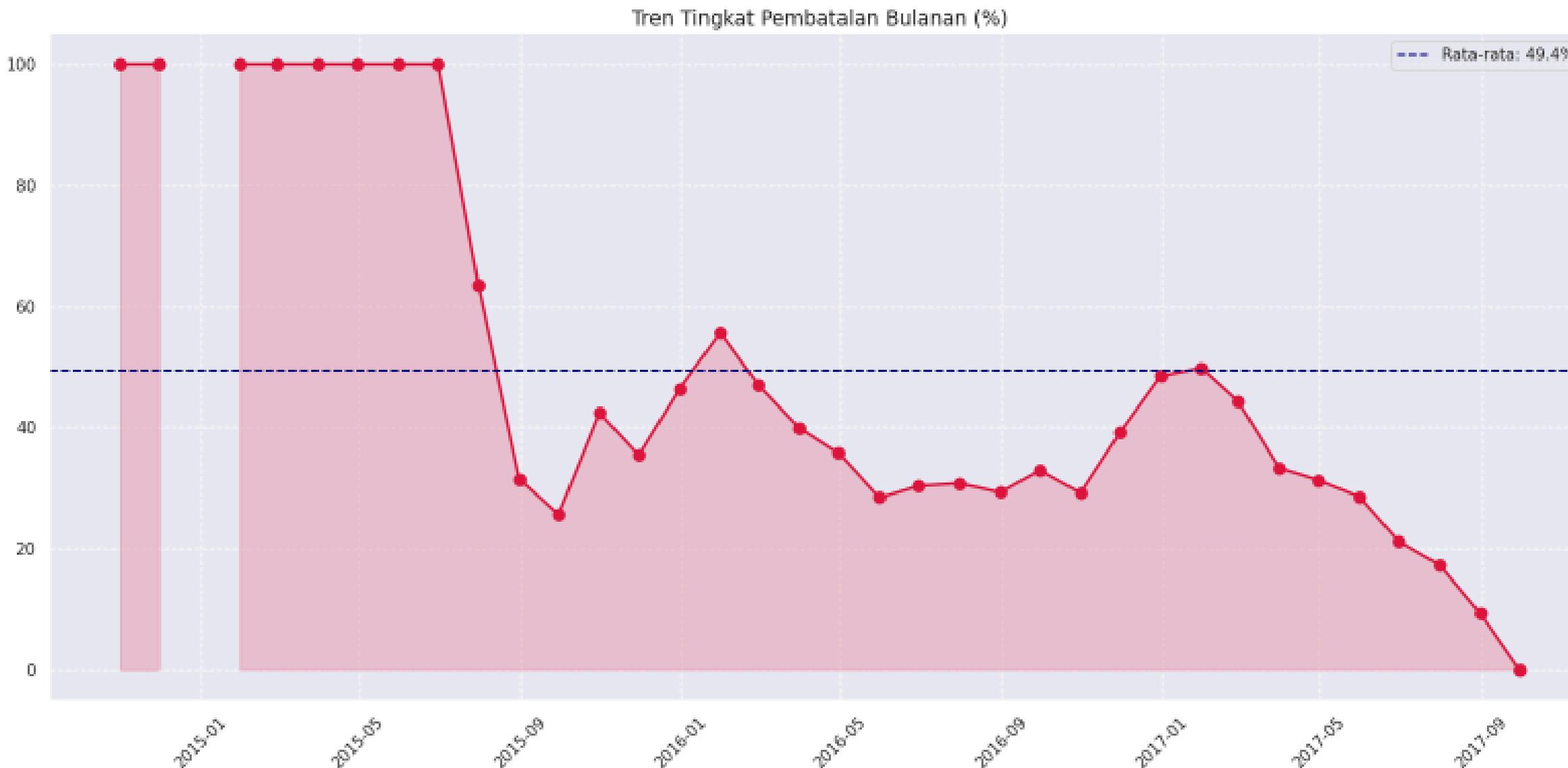
Exploratory Data

Mengatasi Duplicate

Setelah mengidentifikasi adanya 11.050 entri duplikat (21,5% dari total data), langkah penanganan yang dilakukan adalah dengan menghapus seluruh data ganda menggunakan fungsi drop_duplicates(). Proses ini secara otomatis akan mempertahankan hanya satu instance unik dari setiap baris data yang identik, sehingga memastikan tidak ada pengaruh bias dalam analisis lebih lanjut.

```
data = data.drop_duplicates()
```

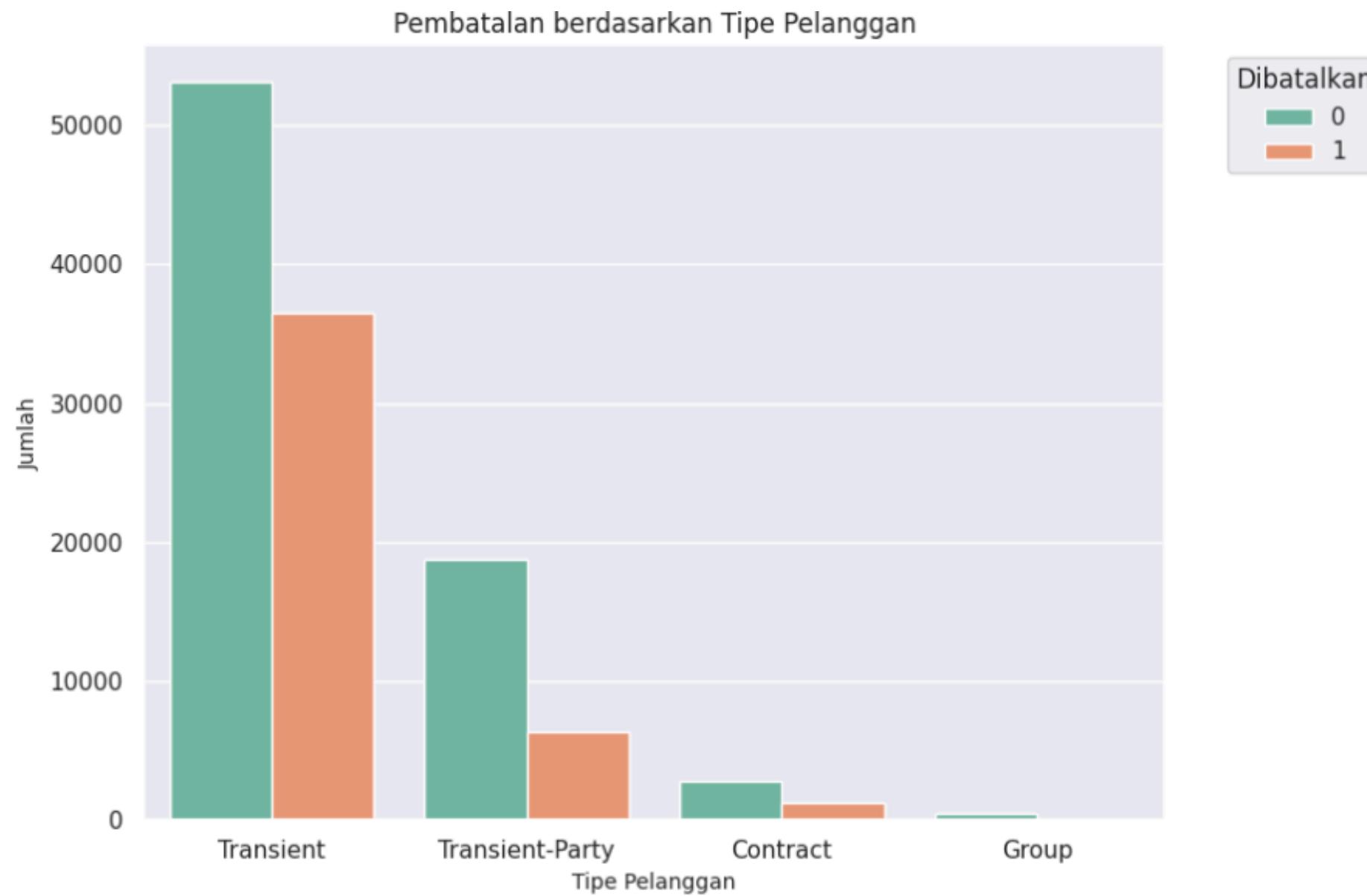
Exploratory Data



Tren Tingkat Pembatalan Bulanan

Grafik menunjukkan tren penurunan tingkat pembatalan pemesanan hotel dari tahun 2015 hingga pertengahan 2017. Pada awal 2015, tingkat pembatalan berada di angka maksimal, yaitu 100%, yang mengindikasikan seluruh pemesanan dibatalkan. Namun, setelah pertengahan 2015, terlihat penurunan yang signifikan dan konsisten, di mana tren mulai stabil di bawah rata-rata keseluruhan sebesar 49,4%. Penurunan ini mencerminkan adanya perbaikan strategi manajemen reservasi, peningkatan kepercayaan pelanggan, atau perubahan dalam segmentasi pasar. Pada akhir periode (2017), tingkat pembatalan bahkan turun sampai 0%, menandakan efisiensi sistem reservasi dan potensi keberhasilan dalam pengelolaan risiko pembatalan.

Exploratory Data

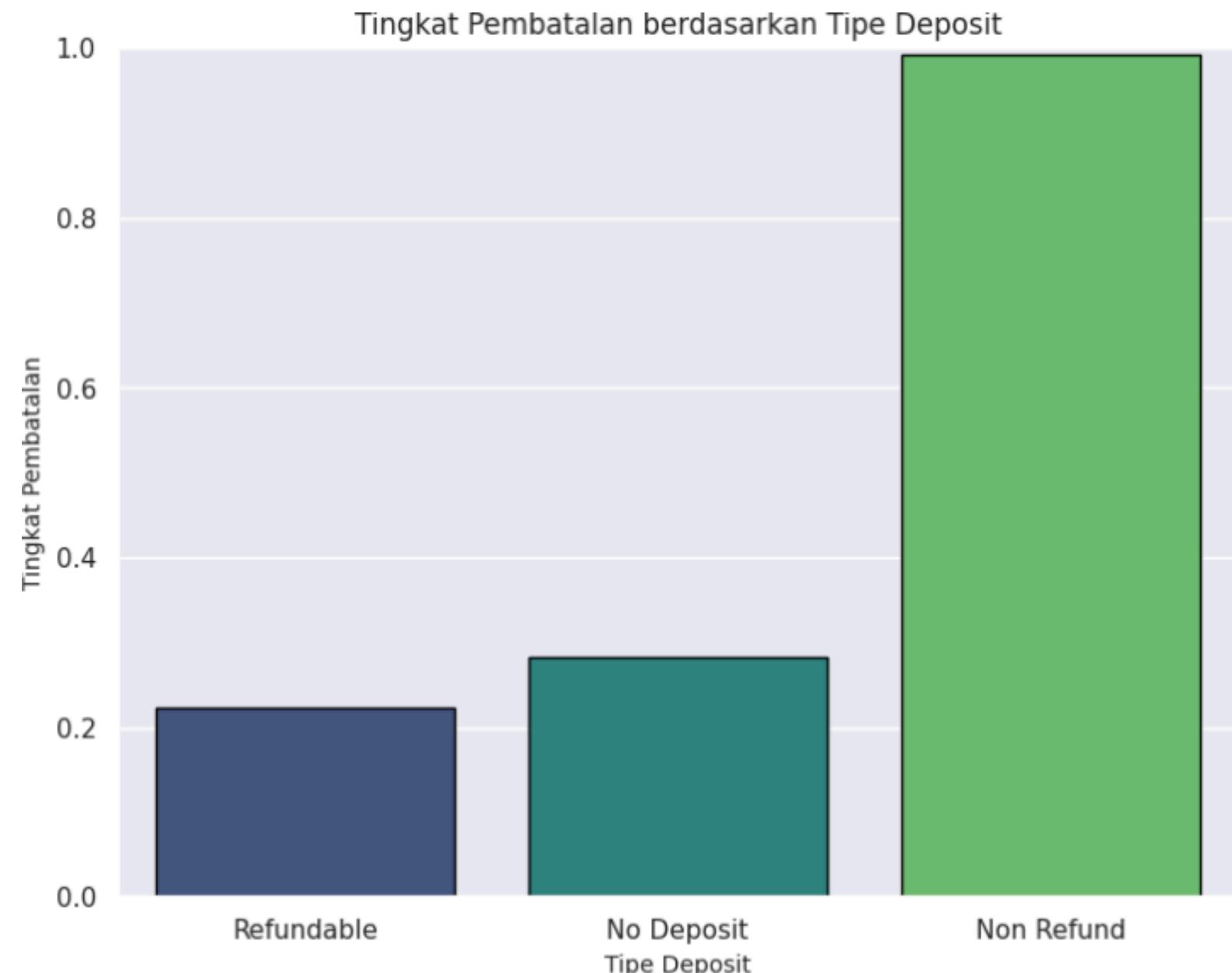


Tren Pembatalan Berdasarkan Tipe Pelanggan

Grafik menunjukkan bahwa pelanggan dengan tipe Transient mendominasi jumlah pembatalan, diikuti oleh Transient-Party, sedangkan tipe Contract dan Group memiliki tingkat pembatalan yang jauh lebih rendah. Hal ini menunjukkan bahwa pelanggan individu yang melakukan pemesanan secara spontan (Transient) cenderung memiliki kemungkinan lebih tinggi untuk membatalkan dibandingkan pelanggan yang melakukan kontrak atau reservasi grup. Strategi mitigasi risiko pembatalan dapat difokuskan pada segmen Transient, misalnya dengan menerapkan kebijakan pembatalan yang lebih ketat atau memberikan insentif untuk mempertahankan reservasi.

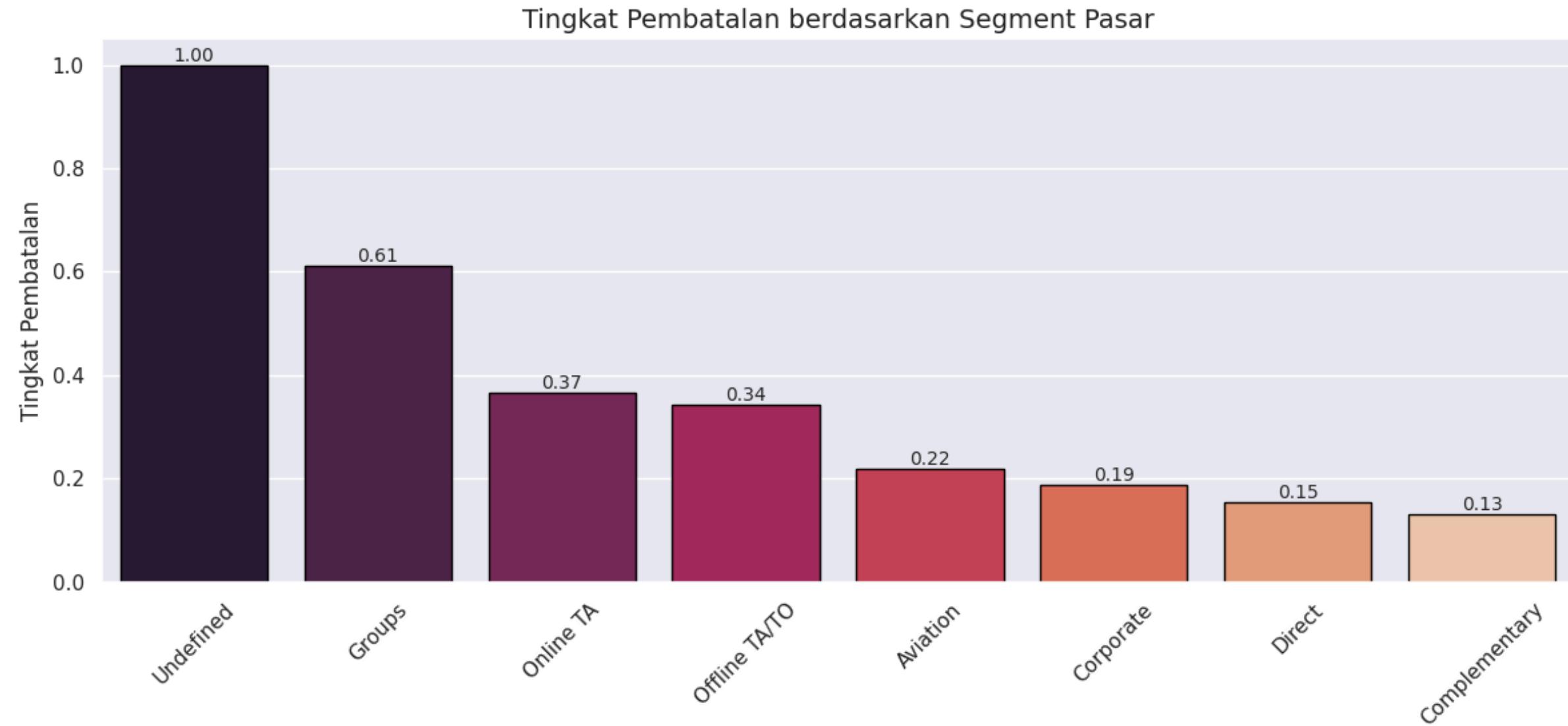
Exploratory Data

Tingkat Pembatalan Berdasarkan Tipe Deposit



Grafik menunjukkan bahwa tingkat pembatalan tertinggi terjadi pada pelanggan dengan tipe deposit Non Refund, yang secara mengejutkan mencapai hampir 100%. Sebaliknya, tingkat pembatalan jauh lebih rendah pada tipe Refundable dan No Deposit, masing-masing di bawah 30%. Hal ini dapat mengindikasikan bahwa meskipun Non Refund seharusnya memberikan disinsentif terhadap pembatalan, kemungkinan besar tipe ini lebih sering digunakan oleh pelanggan dengan risiko tinggi pembatalan. Temuan ini menekankan pentingnya evaluasi ulang terhadap strategi pricing dan kebijakan refund untuk menurunkan risiko pembatalan.

Exploratory Data

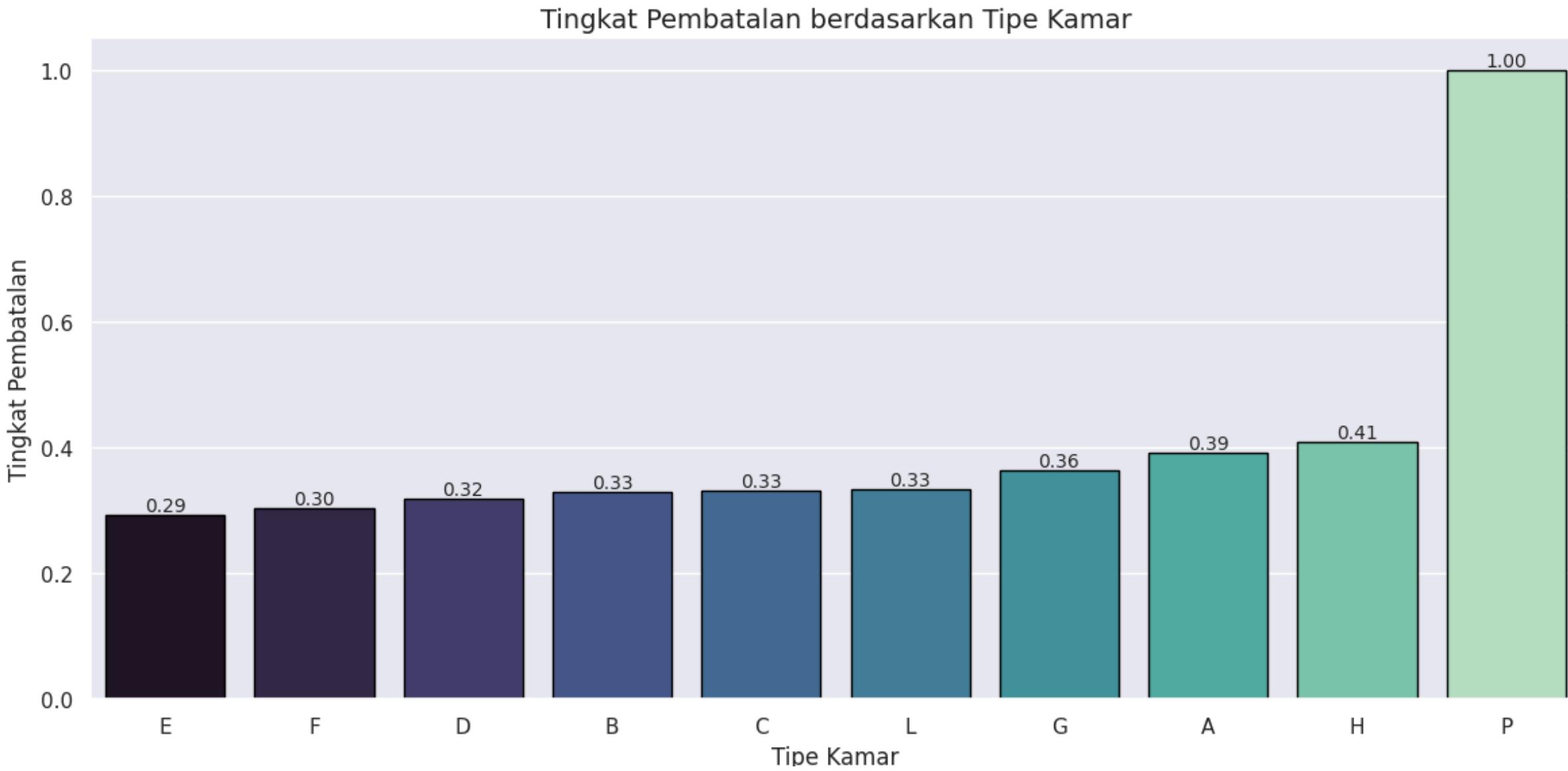


Tingkat Pembatalan Berdasarkan Segment Pasar

Dari grafik terlihat bahwa segmen pasar dengan tingkat pembatalan tertinggi adalah Undefined dengan angka 100%, yang kemungkinan besar disebabkan oleh entri data yang tidak lengkap atau tidak diklasifikasikan dengan benar. Segmen Groups menempati posisi kedua dengan tingkat pembatalan sebesar 61%, yang mengindikasikan bahwa pemesanan kelompok lebih rentan terhadap perubahan rencana. Sebaliknya, segmen-segmen seperti Corporate, Direct, dan Complementary memiliki tingkat pembatalan yang rendah, masing-masing di bawah 20%. Hal ini menunjukkan bahwa pelanggan dari segmen tersebut cenderung lebih pasti dalam melakukan pemesanan. Temuan ini penting sebagai dasar dalam menentukan strategi pemasaran dan kebijakan penanganan pembatalan berdasarkan karakteristik segmen pasar.

Tingkat Pembatalan Berdasarkan Segment Pasar

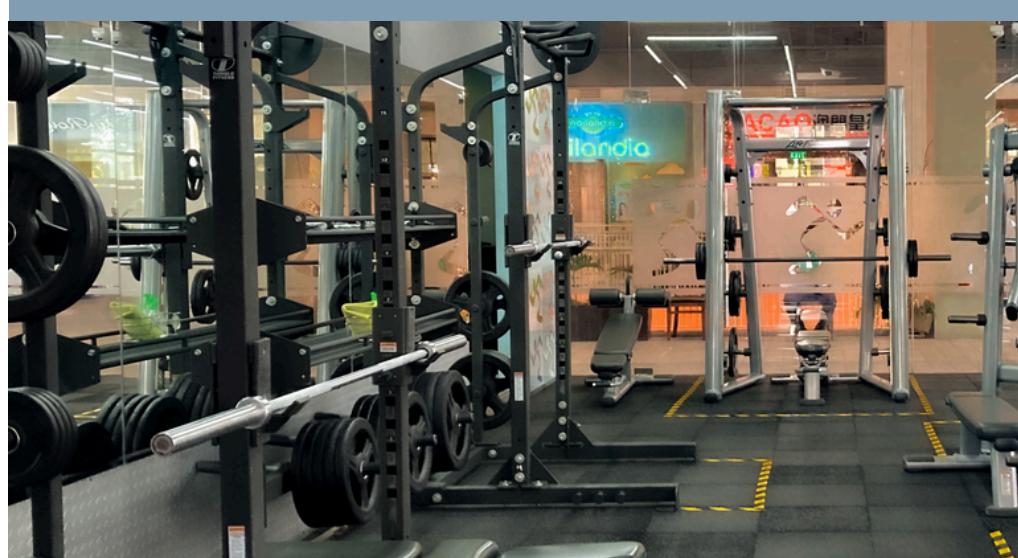
Exploratory Data



Dari grafik terlihat bahwa segmen pasar dengan tingkat pembatalan tertinggi adalah Undefined dengan angka 100%, yang kemungkinan besar disebabkan oleh entri data yang tidak lengkap atau tidak diklasifikasikan dengan benar. Segmen Groups menempati posisi kedua dengan tingkat pembatalan sebesar 61%, yang mengindikasikan bahwa pemesanan kelompok lebih rentan terhadap perubahan rencana. Sebaliknya, segmen-segmen seperti Corporate, Direct, dan Complementary memiliki tingkat pembatalan yang rendah, masing-masing di bawah 20%. Hal ini menunjukkan bahwa pelanggan dari segmen tersebut cenderung lebih pasti dalam melakukan pemesanan. Temuan ini penting sebagai dasar dalam menentukan strategi pemasaran dan kebijakan penanganan pembatalan berdasarkan karakteristik segmen pasar.

Kesimpulan

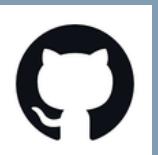
- Kualitas Data: Dataset mengandung missing value dan duplikasi yang cukup signifikan—hingga 21,5% data duplikat dan ribuan nilai kosong. Proses pembersihan data berhasil meningkatkan kualitas dataset untuk analisis lebih lanjut.
- Tren Pembatalan: Terjadi penurunan tren pembatalan secara signifikan dari tahun 2015 ke 2017, yang mencerminkan peningkatan efisiensi sistem reservasi dan kepercayaan pelanggan.
- Faktor Pembatalan: Pembatalan paling tinggi terjadi pada:
 - Segmen pasar Transient dan Groups,
 - Tipe deposit Non Refund, dan
 - Tipe kamar serta market segment Undefined, yang menandakan potensi masalah pada input data atau pelanggan berisiko tinggi.
- Implikasi Strategis: Fokus strategi mitigasi pembatalan sebaiknya diarahkan pada:
 - Segmen Transient dan Group,
 - Evaluasi ulang kebijakan Non Refund,
 - Perbaikan sistem klasifikasi data agar tidak ada entri Undefined.



Terimakasih



[Muhammad Shobri Al Mughdhor](#)



[Shobri24](#)



shobri2424@gmail.com

