# Analisa Prediksi Pembatalan Pemesanan

Studi Kasus customer Hotel

# Deskripsi

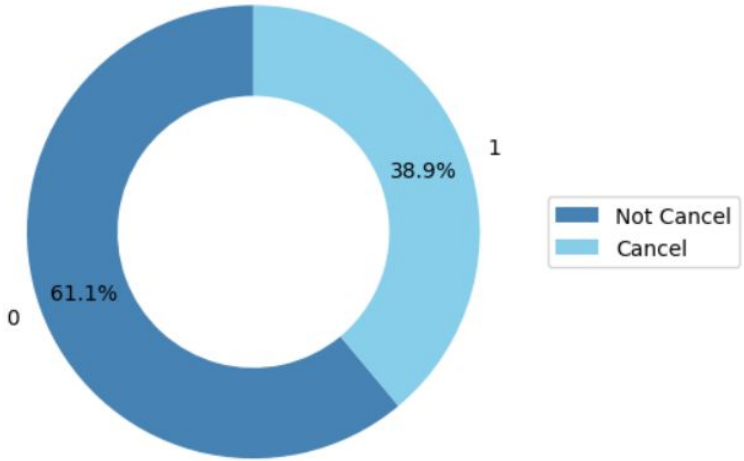
Dataset

Proyek ini menggunakan data train.csv yang terdiri dari 64 hotel yang tersebar di benua Amerika dan Eropa serta pengunjung dari 162 negara. Terdapat 32 kolom dengan 83293 row (pemesanan). Data ini dimulai dari Januari 2017 sampai September 2019.

Tujuan :

- Membangun model prediksi untuk memperkirakan apakah pelanggan akan membatalkan pemesanan.
- Memberikan business insight untuk kepentingan bisnis
- Memberikan rekomendasi bisnis berdasarkan analisa prediksi dan business insight

# Distribusi Cancellation



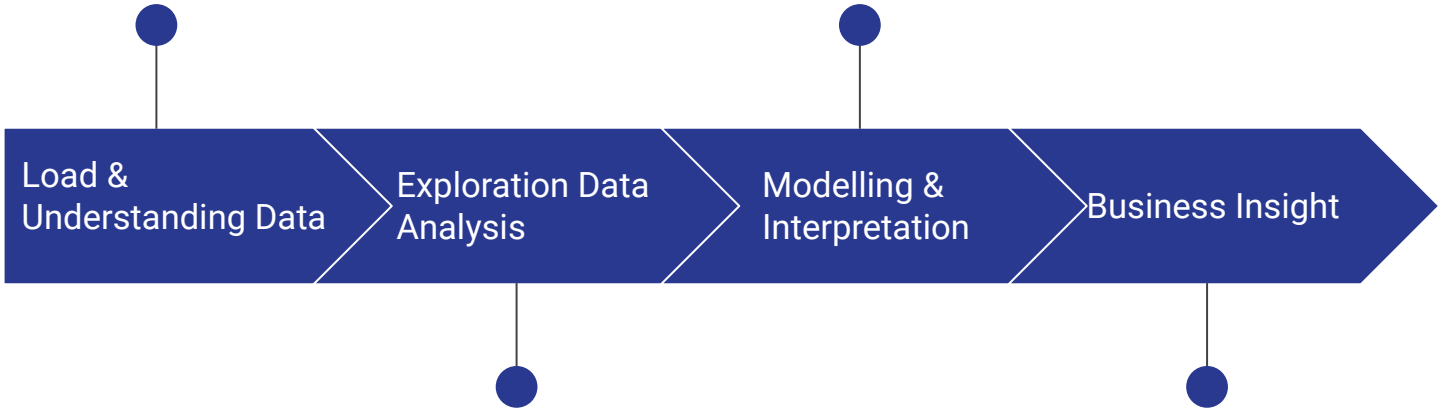

Understanding why Cancellation is important

- Mengurangi biaya karena pembatalan *last-minute* agar memaksimalkan penjualan.

- Memberi insight untuk membuat strategi lebih baik dengan penawaran diskon ke target yang sesuai.

Tahap ini mengumpulkan data serta memahami karakteristik data.

Tahap ini melakukan modelling data dan evaluasi model.




Tahapan dalam EDA :

- Basic Data Cleansing (Missing Data, Duplicate Data)
- Memahami data
- Memilih fitur dlm ML yang akan dibangun

Tahap ini melakukan business question untuk mendapatkan insight serta rekomendasi bisnis.

# Information Dataset

## Basic Pemesanan

- hotel
- is_canceled
- lead_time
- arrival_date_year
- arrival_date_month
- arrival_date_week_number
- arrival_date_day_of_month
- stays_in_weekend_nights
- stays_in_week_nights
- adults
- children
- babies

## Detail Pemesanan

- meal
- country
- market_segment
- distribution_channel
- is_repeated_guest
- previous_cancellations
- previous_bookings_not_canceled

## Detail Akomodasi

- reserved_room_type
- assigned_room_type
- booking_changes
- deposit_type
- agent
- company

## Informasi Tambahan

- days_in_waiting_list
- customer_type
- adr
- required_car_parking_space
- total_of_special_request
- reservation_status
- reservation_status_date

## Handling Missing & Duplicated Values

Terdapat 4 Kolom Null Values
- Company = 94%
- Agent = 13.6%
- Country = 0.4%
- Children = 0.003%

Tidak ada duplicated values

Di drop column > 40% dan drop value < 30% agar tidak bias saat EDA dan pemodelan

## Another Check

- Total Pengunjung < 0 (adults+children+babies) : 83 row
- Total menginap < 0 : 308 row
- 29 Feb : 57 row
- babies > 3 dengan adults 1-2 : 2 row

Drop semua rows

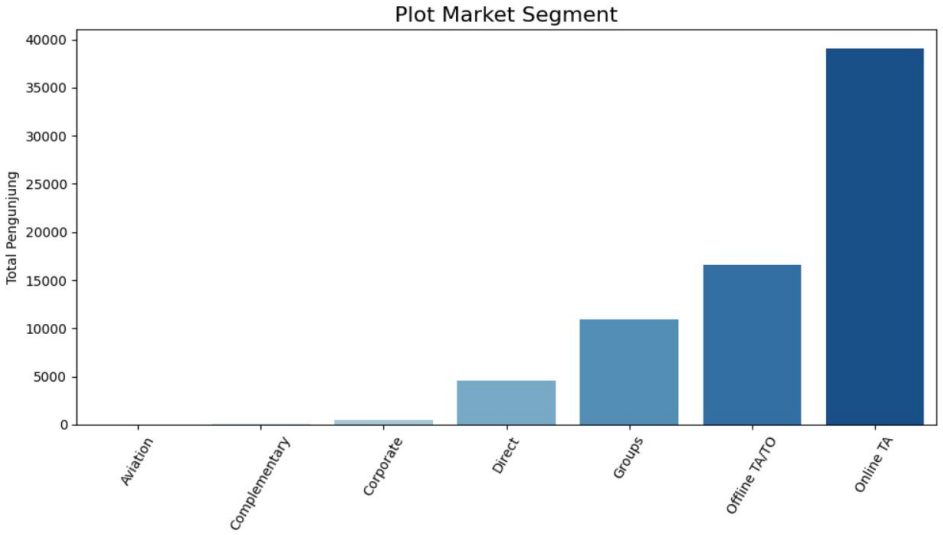
## Outlier Check

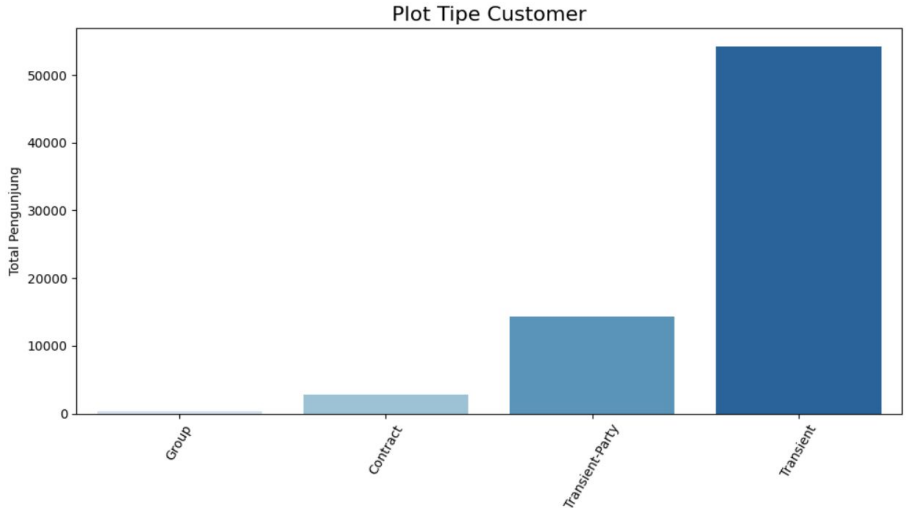
Terdapat banyak outliers, namun hanya ada 1 yang aneh (total bayi > 3) dan yang lain masih dikatakan normal

Drop value babies > 3

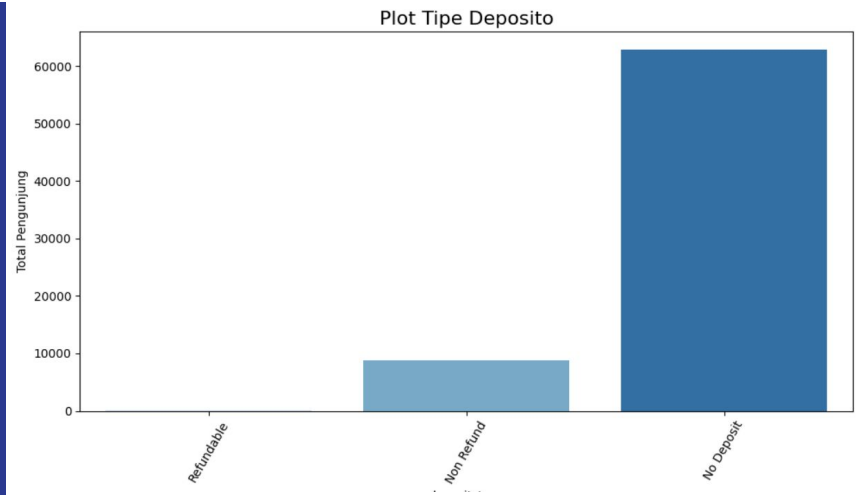
# INFORMASI STATISTIK DESKRIPTIF

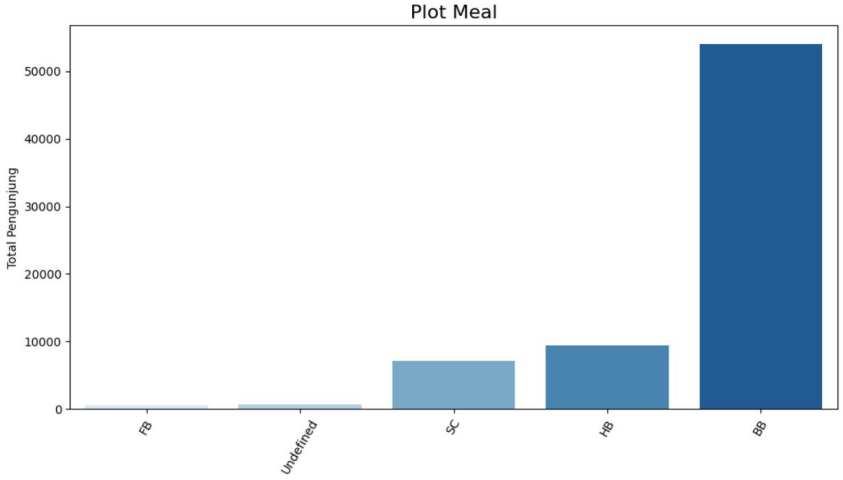
- Rata-rata jarak antara waktu reservasi room dengan kedatangan 111 hari
- Pengunjung biasanya menginap 1 malam pada weekend dan 2 malam pada weekday
- Segmentasi usia pengunjung saat menginap untuk dewasa sebanyak 1-2 orang, dan jarang membawa anak/bayi (rata2 untuk children 1 orang dan bayi bahkan jarang)
- Rata-rata pengunjung masuk kedalam waiting list selama 2 hari dan maksimal 19 hari. Dengan maksimal 19 hari bisa dipastikan terjadi karena *peak season* (musim libur), customer meminta spesifik tipe kamar atau bahkan terjadi perpanjangan menginap oleh tamu yang sedang menginap sehingga untuk tamu yang ingin menginap harus masuk kedalam waiting list.
- Rata-rata ADR  yaitu 104
- Request parkir jarang terjadi karena request terbanyak yaitu 3 request dalam kurun 3 tahun
- Rata-rata spesial request yaitu sebanyak 1 request dan terbanyak sampai 5 request, ini menandakan pengunjung jarang menggunakan special request.

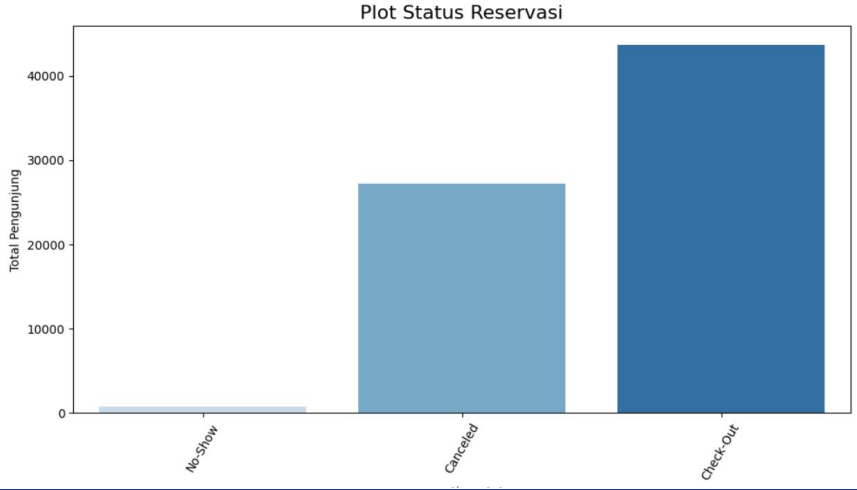


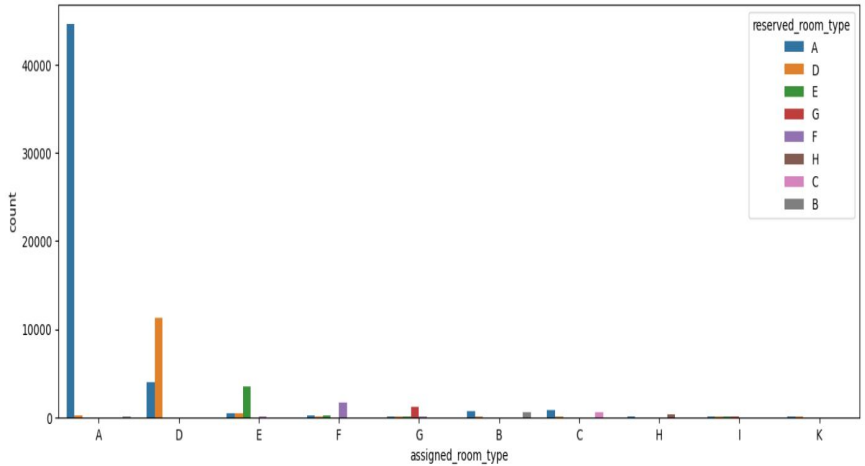




- Segment pasar terbanyak yaitu dari agen travel online (Online TA), agen travel secara offline (Offline) dan rombongan tur/keluarga besar dan hanya sedikit yang memesan hotel untuk market customer keperluan bisnis (corporate)
- Kebanyakan customer memesan untuk jangka pendek 1/2 malam (Transient) dan sekelompok customer yang tinggal dalam jangka pendek dan hanya sedikit untuk kontrak (urusan bisnis)
- Kebanyakan customer memesan hotel tanpa membayar (No Desposit) dibandingkan customer yang membayar uang muka (Non Refund) atau Refundable






- Customer yang checkout mendominasi namun customer yang membatalkan pesanan cukup banyak dan hanya sedikit customer yang meng-ghosting reservasinya.
- BB(Bed and Breakfast) paling umum/diminati customer dan diikuti oleh HB (Half Board) dan hanya sedikit customer yang full mengambil paket hotel (Full Board).
- Tipe kamar yang awal dipesan paling banyak yaitu tipe A, D dan E dan saat assigned frekuensinya menurun, hal ini kemungkinan terjadi ada perubahan pemesanan saat datang ke hotel.

# Modelling

Baseline Model, With Hyperparameter Tunning and Oversampling

# Baseline Model

| Model | Recall | AUC | F1 Score | Precision | Accuracy |
|---|---|---|---|---|---|
| LogisticRegression | 0.534306 | 0.717355 | 0.632115 | 0.773758 | 0.757876 |
| DecisionTreeClassifier | 0.686350 | 0.772040 | 0.718873 | 0.754633 | 0.791009 |
| RandomForestClassifier | 0.710256 | 0.781555 | 0.731818 | 0.754731 | 0.797338 |
| XGBClassifier | 0.666627 | 0.779134 | 0.725936 | 0.796828 | 0.804039 |

# Hyperparameter Tunning

| Model | Recall | AUC | F1 Score | Precision | Accuracy |
|---|---|---|---|---|---|
| LogisticRegression | 0.533947 | 0.717518 | 0.632227 | 0.774848 | 0.758155 |
| DecisionTreeClassifier | 0.570284 | 0.723419 | 0.646608 | 0.746519 | 0.757318 |
| RandomForestClassifier | 0.446569 | 0.704349 | 0.593063 | 0.882589 | 0.761413 |
| XGBClassifier | 0.660053 | 0.773942 | 0.719010 | 0.789534 | 0.799153 |

# Oversampling - SMOTE

| Model | Recall | AUC | F1 Score | Precision | Accuracy |
|---|---|---|---|---|---|
| LogisticRegression | 0.658857 | 0.729985 | 0.668607 | 0.678651 | 0.745730 |
| DecisionTreeClassifier | 0.743366 | 0.778374 | 0.730187 | 0.717467 | 0.786123 |
| RandomForestClassifier | 0.761774 | 0.785901 | 0.739670 | 0.718813 | 0.791242 |
| XGBClassifier | 0.771814 | 0.789778 | 0.744494 | 0.719042 | 0.793755 |

# Best Model

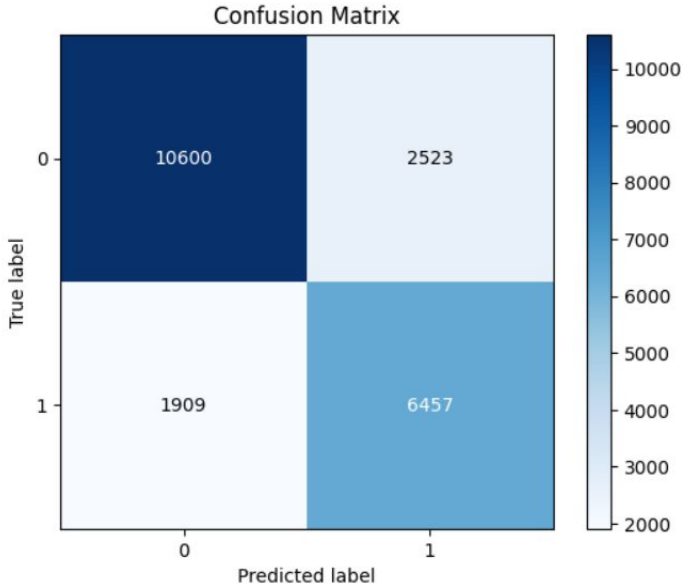
Untuk menilai performa model disini menggunakan evaluasi metrik yaitu Precision, Recall/Sensitivitas dan AUC.

- AUC : mampu membedakan kelas positif negatif, namun walaupun hasilnya bagus lebih baik dikombinasi dengan presisi dan recall.
- Presisi : fokus pada hasil prediksi positif kita, jadi dari hasil prediksi 100 orang berapa yang benar2 positif itu presisi.
- Recal : untuk mengukur sensitivitas performa model baik itu dalam keadaan balance maupun imbalance, fokus pada hasil aktualnya, dari 100 orang yg beneran cancer maka model kita berhasil menangkap 80%.

| Model | Recall | AUC | F1 Score | Precision | Accuracy |
|---|---|---|---|---|---|
| LogisticRegression | 0.658857 | 0.729985 | 0.668607 | 0.678651 | 0.745730 |
| DecisionTreeClassifier | 0.743366 | 0.778374 | 0.730187 | 0.717467 | 0.786123 |
| RandomForestClassifier | 0.761774 | 0.785901 | 0.739670 | 0.718813 | 0.791242 |
| XGBClassifier | 0.771814 | 0.789778 | 0.744494 | 0.719042 | 0.793755 |

Model terbaik yaitu XGBoost Classifier dimana :

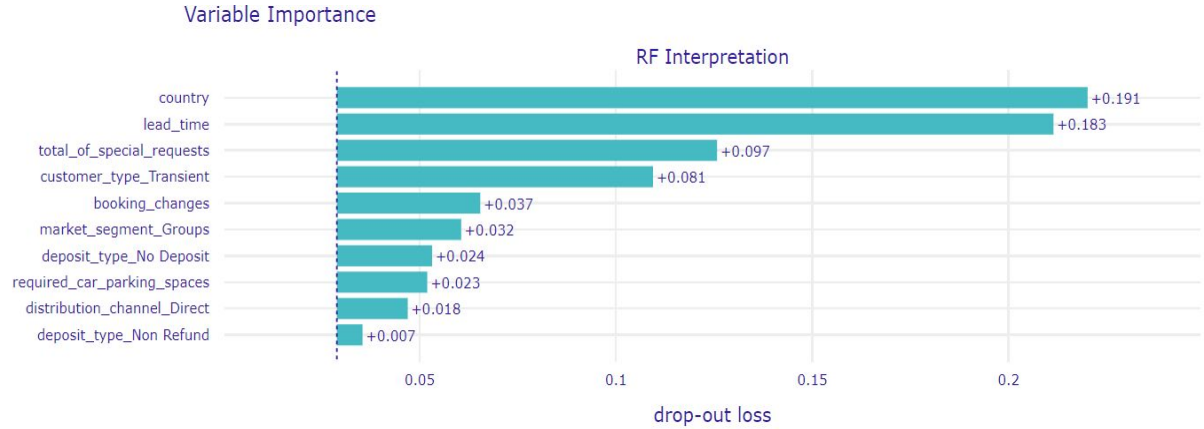
- Memiliki nilai Akurasi dan AUC tertinggi menandakan performa model bekerja dengan baik dalam membedakan data positif negatif.
- Memiliki nilai Recall yang baik yaitu dari 100 orang yang ***beneran*** cancel maka model kita berhasil menangkap sebesar 77% yang benar2 cancel.
- Presisi 72% yaitu dari 100 orang yang ***diprediksi*** cancel maka model menangkap 72% orang yang betul2 cancel




# XGBoost Classifier

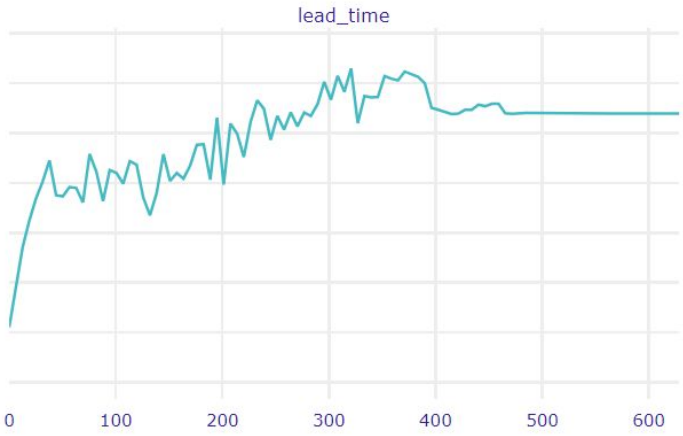
Memiliki nilai Recall yang baik yaitu dari 100 orang yang ***beneran*** cancel maka model kita berhasil menangkap sebesar 77% yang benar2 cancel. Presisi 72% yaitu dari 100 orang yang ***diprediksi*** cancel maka model menangkap 72% orang yang betul2 cancel

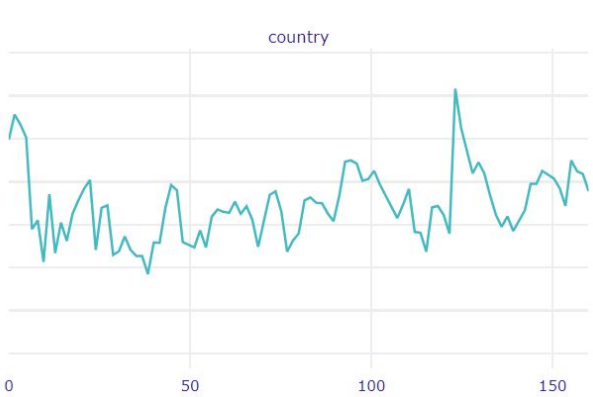
Yang diprediksi cancel namun sebenarnya tidak yaitu 2523 customer tetapi yang diprediksi tidak cancel namun aslinya cancel sebanyak 1909 customer



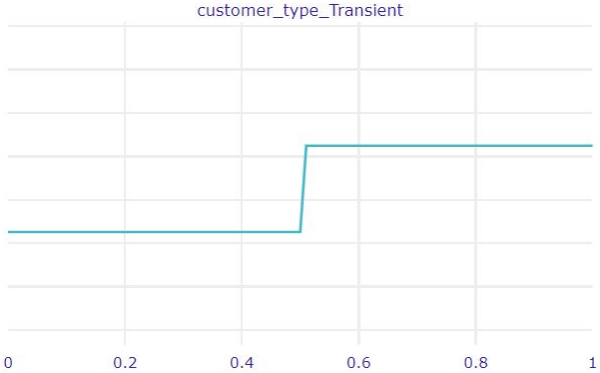

Variabel yang mempengaruhi cancel tidaknya yaitu negara, jarak waktu reservasi dan kedatangan, special request, tipe customer yang tinggal jangka pendek ½ hari dan booking changes.

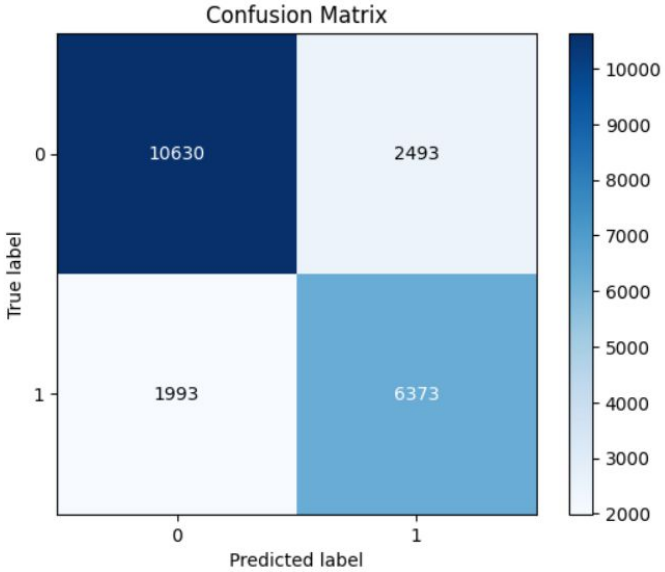

total_of_special_requests
prediction
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5


lead_time
0
100
200
300
400
500
600


country
0
50
100
150


booking_changes
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5
10
15

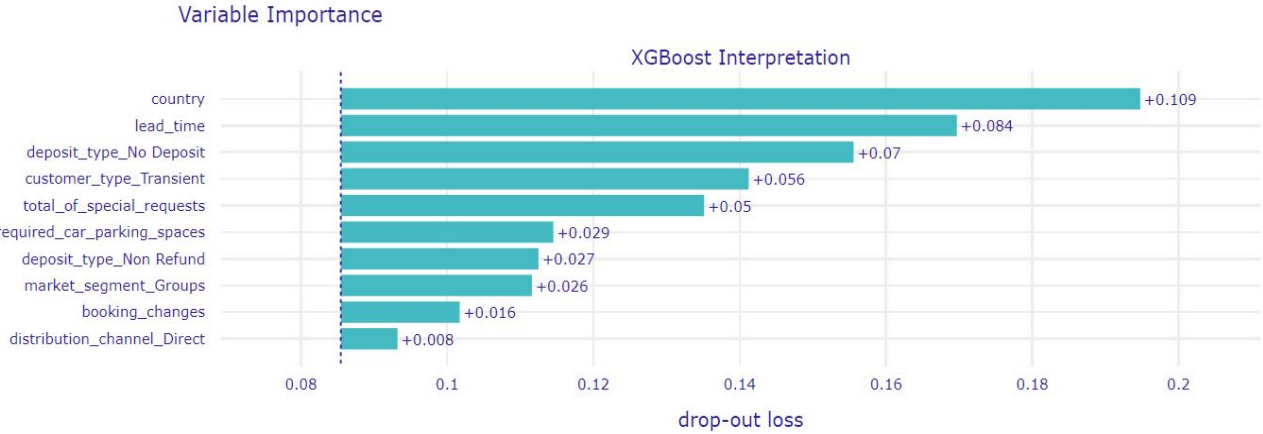

customer_type_Transient
0
0.2
0.4
0.6
0.8
1



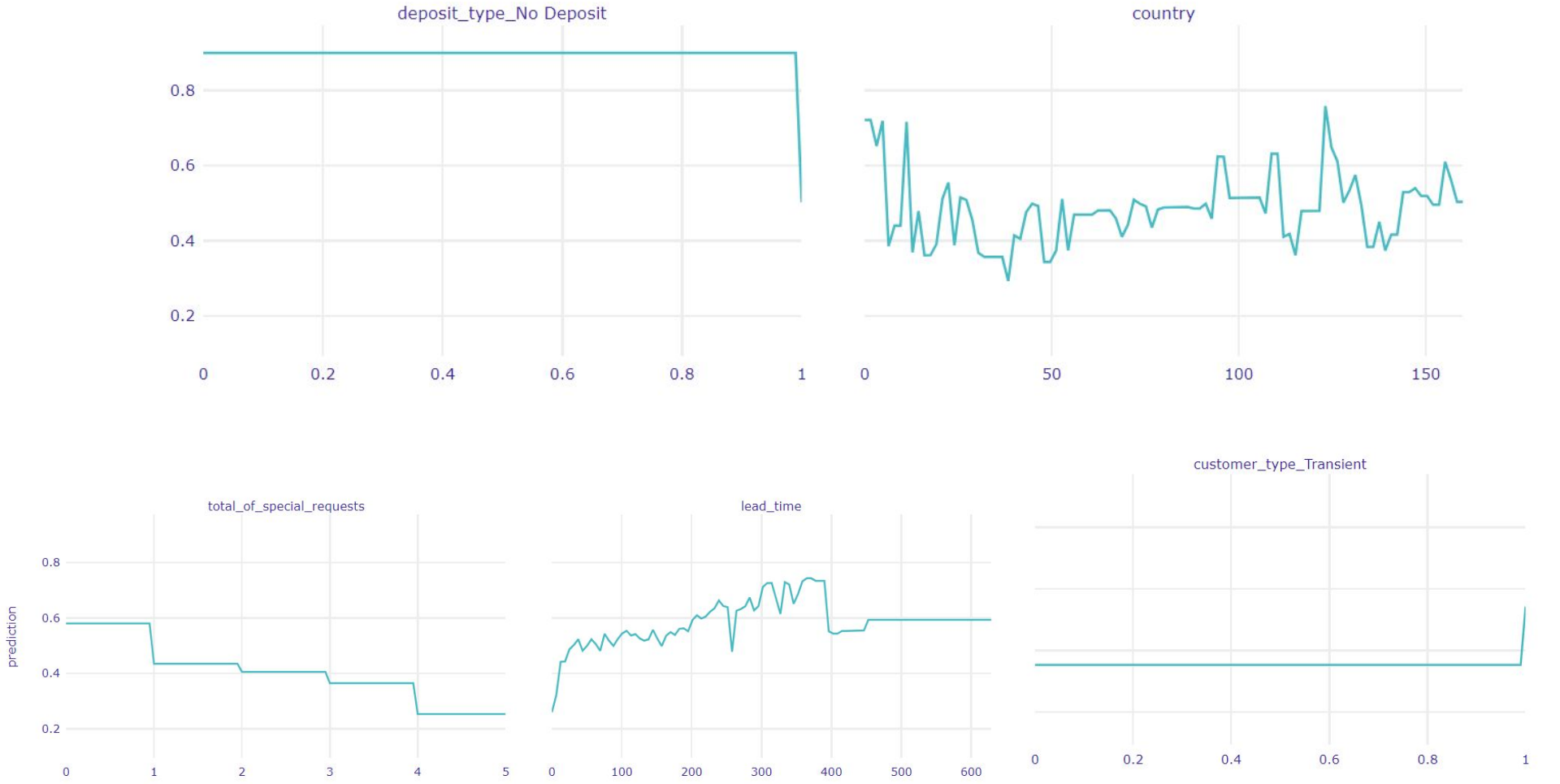

# Random Forest

memiliki recall 76% yaitu dari 100 orang yang ***beneran*** cancel maka model kita berhasil menangkap sebesar 76% yang benar2 cancel dan presisi 72% yaitu dari 100 orang yang ***diprediksi*** cancel maka model memprediksi 76% orang betul2 cancel.

Yang diprediksi cancel namun sebenarnya tidak yaitu 2493 customer tetapi yang diprediksi tidak cancel namun aslinya cancel sebanyak 1993 customer

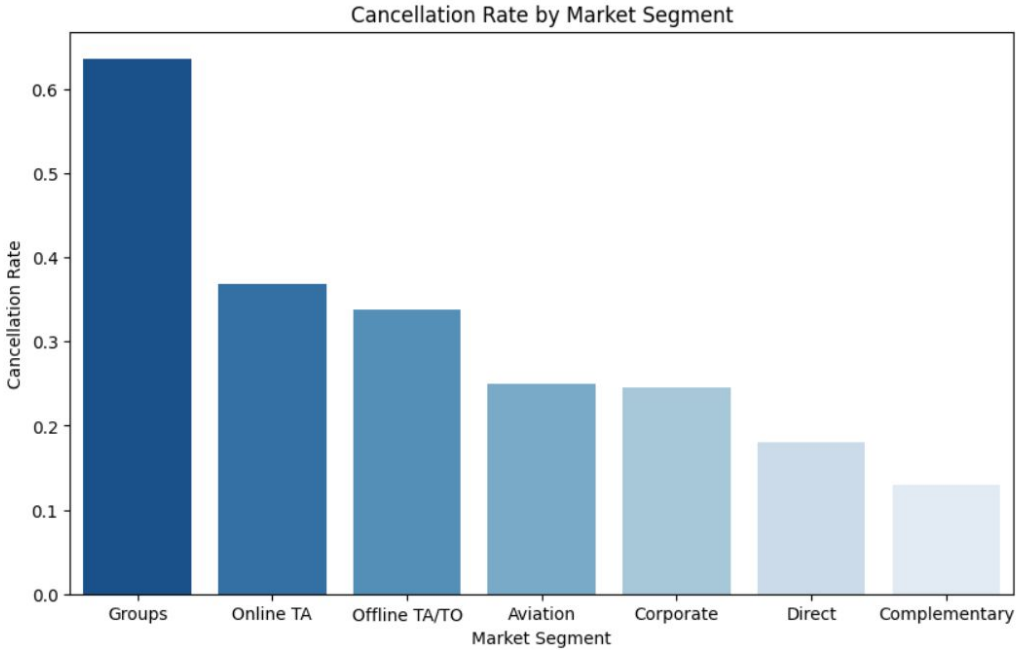



Variabel yang mempengaruhi cancel tidaknya yaitu negara, jarak waktu reservasi dan kedatangan, tipe deposit tanpa uang muka, tipe customer yang tinggal jangka pendek ½ hari, special request, dan request parkir.


deposit_type_No Deposit
country
total_of_special_requests
lead_time
customer_type_Transient
prediction
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1
50
100
150
1
2
3
4
5
200
300
400
500
600

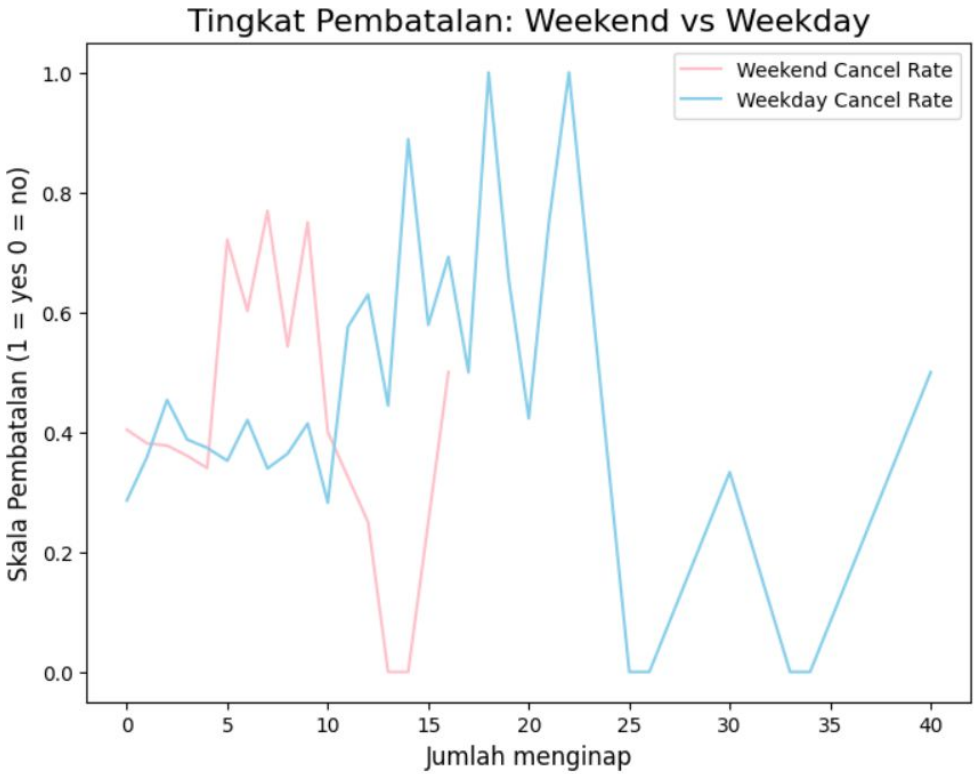
# Business Question

# Market penjualan yang paling tinggi persentase Cancelnya?




Segment market terbanyak dari travel online maupun offline diikuti dengan rombongan/Group. Dilihat dari grafik disamping, tingkat cancel tertinggi yaitu pada Group (rombongan), untuk travel online/offline termasuk lumayan tinggi dan untuk penerbangan, corporate memiliki rata-rata cancel yang rendah.

# Pada hari apakah yang tinggi persentase Cancelnya?

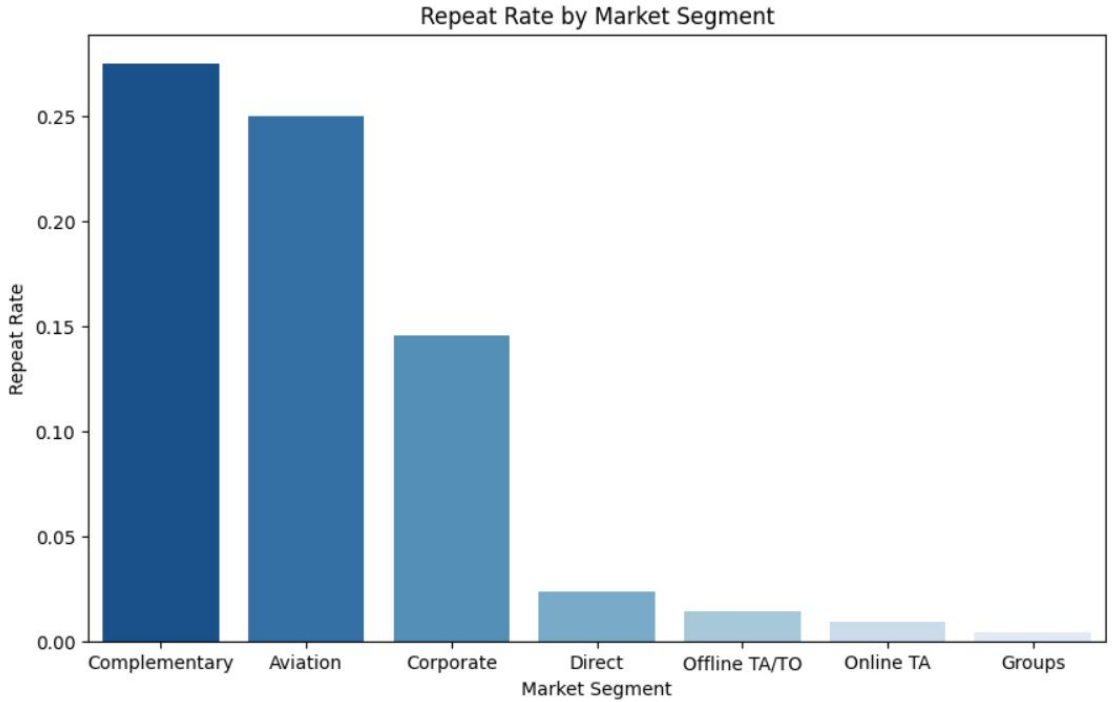



Tingkat cancel paling tinggi saat hari biasa (weekday) dengan durasi waktu menginap 12 - 23 hari. sedangkan akhir pekan memiliki tingkat cancel yang relatif rendah dengan durasi waktu menginap 3 - 8 hari.

# Berapa persen pengunjung yang melakukan repeat stay?

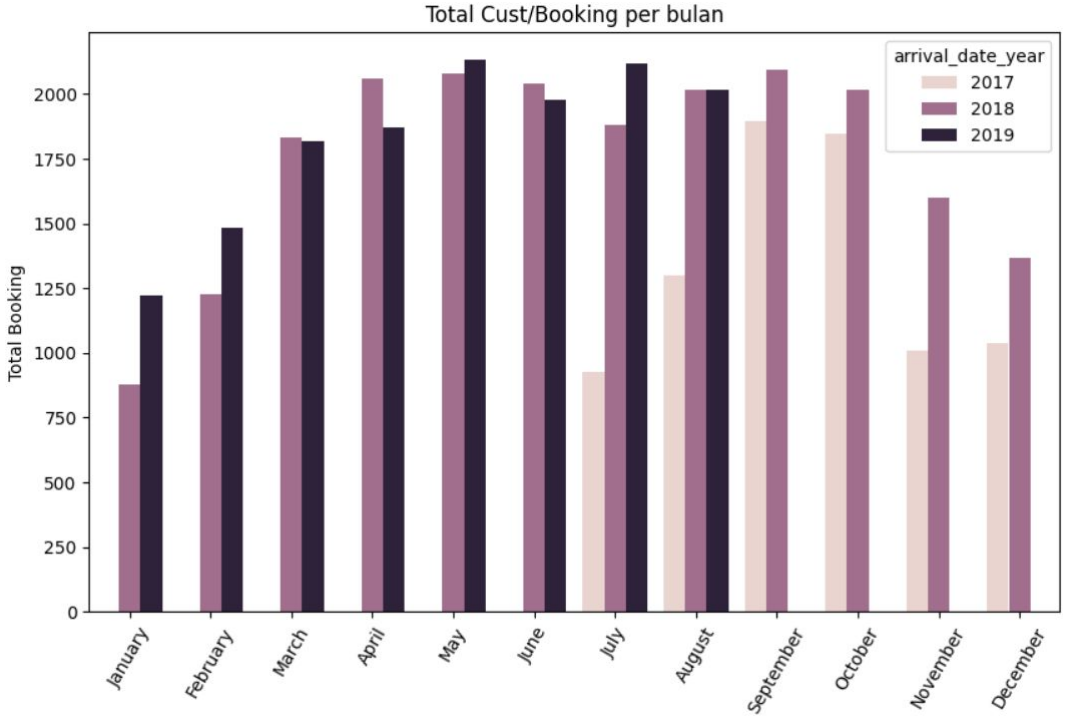

Jika dilihat dari total pengunjung dari 164 negara di beberapa hotel hanya 1.2% pengunjung yang melakukan repeat stay.

# Segment market apa yang paling tinggi melakukan repeat stay?




Dari 1.2% yang melakukan repeat stay, segmentasi nya yaitu berasal dari hadiah/bonus kantor (complementary), Penerbangan (aviation) dan keperluan bisnis (corporate)

# Berapa jumlah pengunjung yang checkout (no cancel) sepanjang 3 tahun ini?




Di tahun 2017 dan 2018 frekuensi pengunjung meningkat di musim gugur (libur musim gugur).

Di tahun 2018 dan 2019 frekuensi pengunjung meningkat, ini terjadi bertepatan dengan musim semi

# Berapa lama rata-rata pengunjung menginap, dari negara mana?

| | country | rata_total_bermalam |
|---|---|---|
| **52** | FRO | 12.000000 |
| **1** | AGO | 9.817259 |
| **131** | SEN | 8.125000 |
| **142** | TGO | 8.000000 |
| **59** | GNB | 7.500000 |
| **56** | GHA | 7.333333 |
| **121** | PLW | 7.000000 |
| **133** | SLE | 7.000000 |
| **98** | MKD | 6.000000 |
| **104** | MRT | 6.000000 |

Rata-rata tertinggi pengunjung menginap berdasarkan asal negara yaitu Denmark dengan lama bermalam 12 hari, Angola dengan 9 hari bermalam dan Senegal serta Togo dengan 8 hari bermalam.

# Output

Jumlah cancel : 40% Total
bermalam rata-rata 1-3 hari
ADR = 104 Dollar
Rata-rata waiting list 2 hari
Total pengunjung/hotel/hari : 2

# Rekomendasi

- Frekuensi pengunjung meningkat di musim semi dan gugur, hal ini bisa dilakukan strategi agar tidak terjadi penumpukan waiting list bahkan cancel
- Jarang terjadi special request ini bisa dianalisa lagi dengan form kepuasan pelanggan dikarenakan dengan adanya special request menunjukan customer memiliki tingkat kepercayaan tinggi kepada hotel dan hal ini bisa meningkatkan kualitas pelayanan serta fasilitas.
- Untuk waktu *peak season*, pihak hotel bisa membuat konfirmasi kedatangan agar tidak terjadinya penumpukan pelanggan dan meminimalisir tingkat cancel.
- Bisa membuat promosi sesuai tipe kunjungan dan frekuensi pelanggan, hal ini terlihat banyaknya pelanggan yang jarang stay lebih dari 2 hari dihari kerja kemungkinan untuk kepentingan bisnis.
- Pertumbuhan reservasi di tahun 2018-2019 sekitar 39% dan pertumbuhan pendapatan berdasarkan ADR sekitar 41%
- Bisa membuat kebijakan jika pemesanan kamar kurang dari 3 bulan bisa *reschedule* atau melakukan pembatalan 10 hari sebelum hari H namun pembatalan atau *reschedule* dilakukan setelah 10 hari sebelum kedatangan maka dikenakan biaya.
- Jika membuat pesanan kamar diatas 3 bulan maka tidak bisa melakukan reschedule dan pembatalan sehingga customer harus melakukan check-in atau melakukan pembayaran ketidakhadiran.

# Dampak

peningkatan penjualan > 41% setelah melakukan pemodelan dan mengimplementasikan strategi pemasaran hasil rekomendasi.