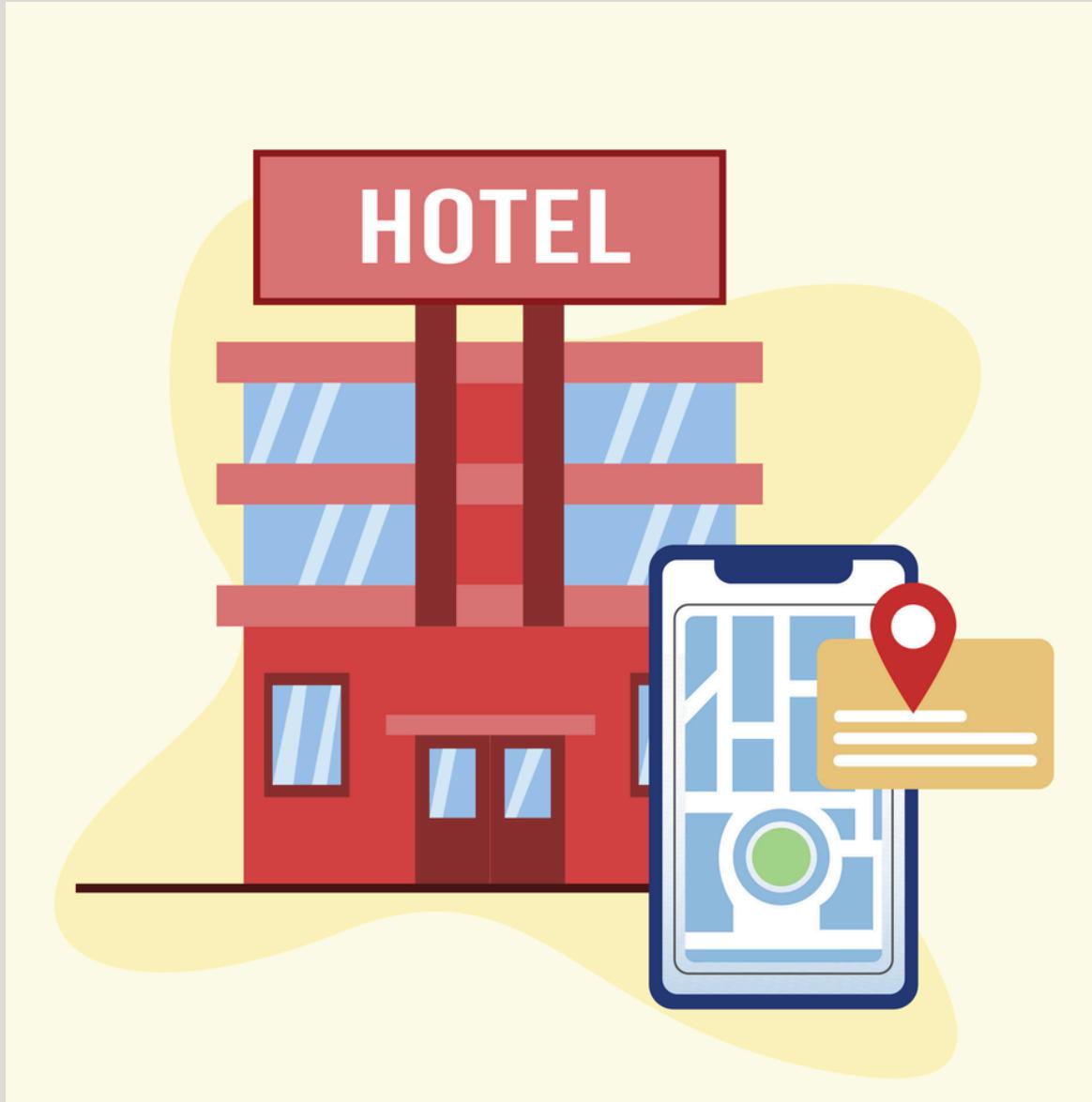


Hotel Booking Study Case



Arya Wiratmaji

[Project Link](#)

Outline

Data Understanding

Data Cleaning & Preprocessing

Exploratory Data Analysis

Model Evaluations

Feature Importance + Analysis



— Data Understanding

Tentang data yang digunakan

Dataset yang digunakan merupakan dataset mengenai informasi pemesanan hotel. Dataset ini memiliki berbagai kolom yang berkaitan dengan perilaku pemesanan, pembatalan, dan metrik relevan lainnya.



Background

Asosiasi perhotelan ingin mengetahui aspek apa saja yang berpengaruh positif ataupun negatif terhadap kesuksesan pemesanan hotel, dan memprediksi kemungkinan pembatalan pemesanan

Goal

Menganalisa aspek yang berpengaruh terhadap kesuksesan pemesanan, dan membuat model prediksi kemungkinan pembatalan pemesanan hotel.

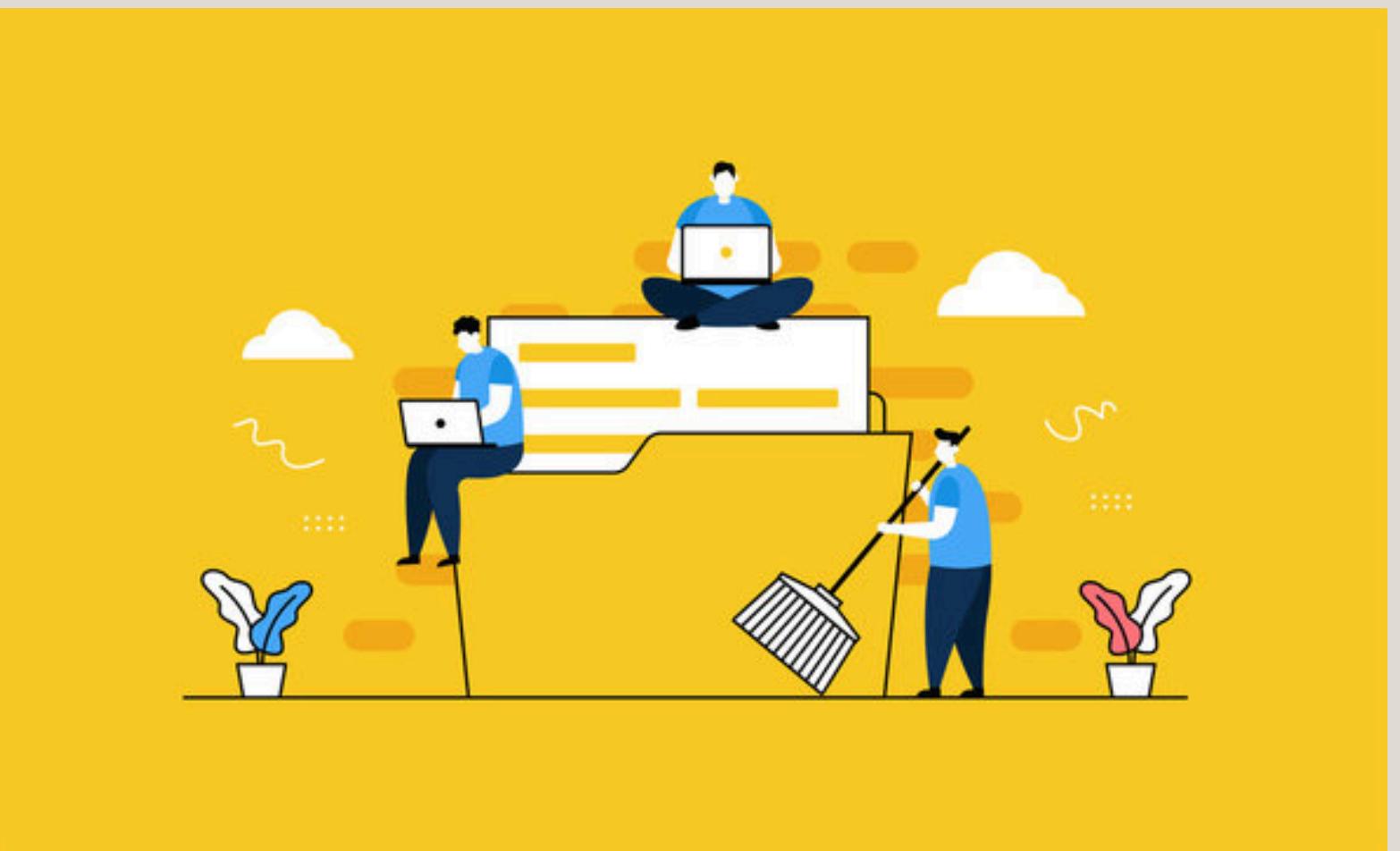
Problems

Asosiasi perhotelan ingin meningkatkan penjualan, namun tidak mengetahui aspek apa saja yang penting terhadap pembatalan pemesanan

Objective

Menganalisa feature untuk menggali informasi lebih dalam & membuat model prediksi untuk memprediksi data.

Data Cleaning & Preprocessing



Missing Values

Train Data

```
children      3  
country     346  
agent      11404  
company    78559  
dtype: int64
```

Missing value percentage	
children	0.003602
country	0.415401
agent	13.691427
company	94.316449

Kolom 'children':

- Train Data = 3 unit | 0.003%
- Test Data = 1 unit | 0.003%

Kolom 'country':

- Train Data = 346 unit | 0.4%
- Test Data = 142 unit | 0.39%

Test Data

```
children      1  
country     142  
agent      4925  
company    33636  
dtype: int64
```

Missing value percentage	
children	0.002801
country	0.397781
agent	13.796291
company	94.223766

Kolom 'agent':

- Train Data = 11.404 unit | 13.7%
- Test Data = 4.925 unit | 13.8%

Kolom 'company':

- Train Data = 78.559 unit | 94.3%
- Test Data = 33.636 unit | 94.2%



Keputusan

‘company’ memiliki NaN > 50%, kita putuskan untuk **drop** entire column.

‘children’ memiliki NaN yang relatif sedikit, dilakukan **imputasi Median**.

‘country’ memiliki NaN yang relatif sedikit.

Setelah dicek apakah ada negara yang mendominasi ternyata ada.

Kita putuskan untuk **imputasi Modus**, karena tidak terlalu mengotori data karena ada negara yang dominan.

‘agent’ karena persentase NaN > 10%, agak risikan jika kita melakukan imputasi, karena bisa mengotori data. (Notes: Value kosong bisa jadi booking yang dilakukan tidak melalui agent)



	country	count	percentage
0	PRT	33952	40.762129
1	GBR	8381	10.062070
2	FRA	7243	8.695809
3	ESP	5897	7.079827
4	DEU	5107	6.131368
..
160	MRT	1	0.001201
161	BRB	1	0.001201
162	NPL	1	0.001201
163	DJI	1	0.001201
164	NAM	1	0.001201

[165 rows x 3 columns]

Pic: Negara yang mendominasi

Setelah diperiksa, ternyata memang tidak ada nilai 0 pada kolom ‘agent’. Berarti NaN bisa jadi adalah booking tanpa agent.



Kita putuskan untuk isi nilai 0 untuk NaN

Duplicate Values

Train Data

Total duplicate train data: 0

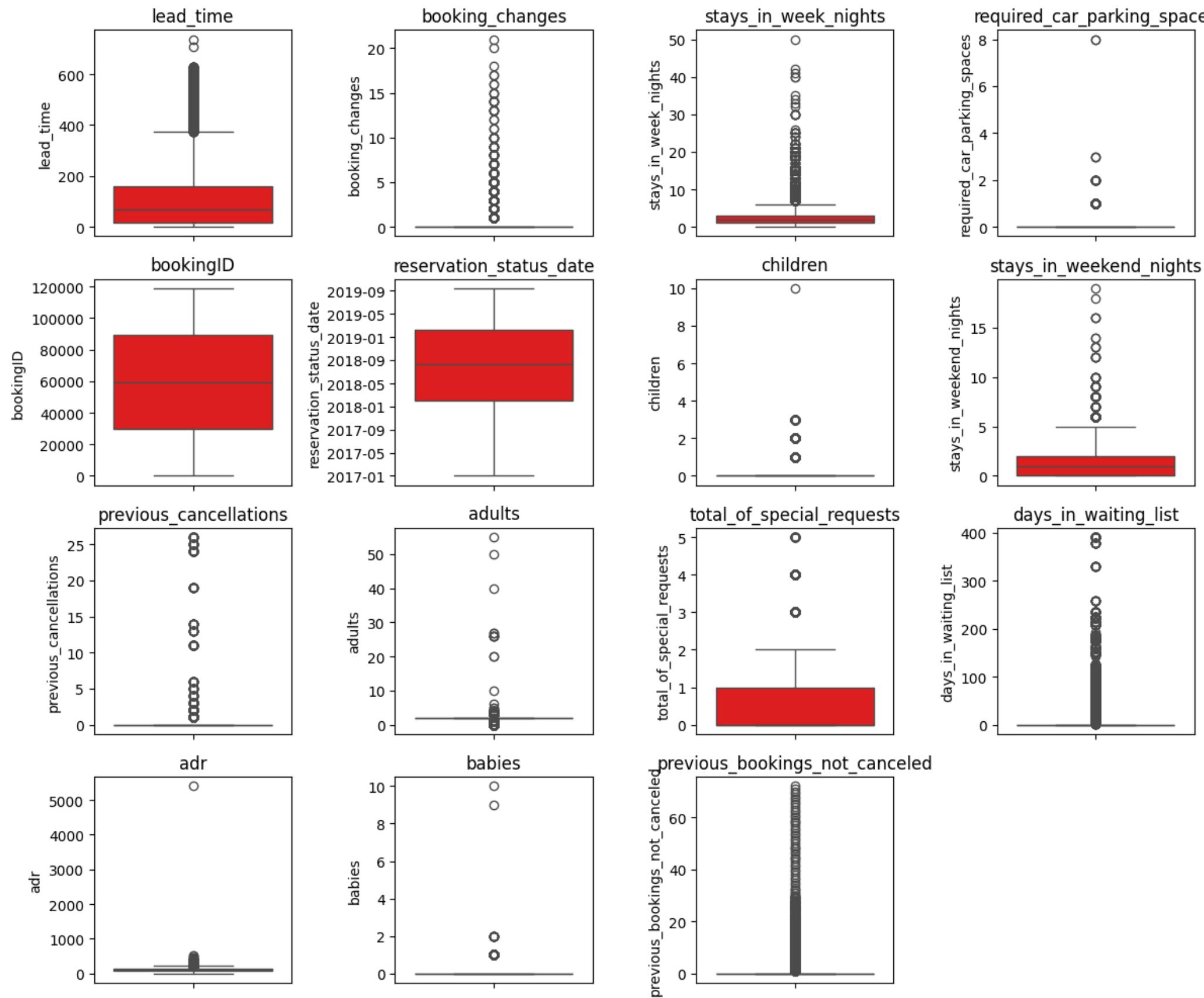
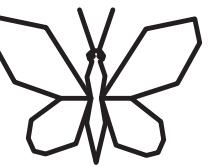
Tidak ada duplicate row berdasarkan aspek pengecekan seluruh kolom

Test Data

Total duplicate test data: 0



Outliers Check



Terdapat outliers pada dataset, namun masih dalam batas wajar dan masuk akal.

Namun terdapat 1 outlier yang nilainya tidak masuk akal

- `stays_in_weekend_nights` = 0
- `stays_in_week_nights` = 1
- ADR = 5400

```
df_outlier_check[df_outlier_check['adr'] > 5000]
```

```
stays_in_weekend_nights stays_in_week_nights a
```

adr

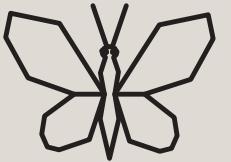
5400.0

Hampir tidak mungkin pendapatan rata-rata per hari pada 1 booking dengan nilai `stays_in_weekend_nights` = 0 dan `stays_in_week_night` = 1 bernilai 5400.

Decision -> Drop



Feature Engineering



Total Nights -> Jumlah malam menginap

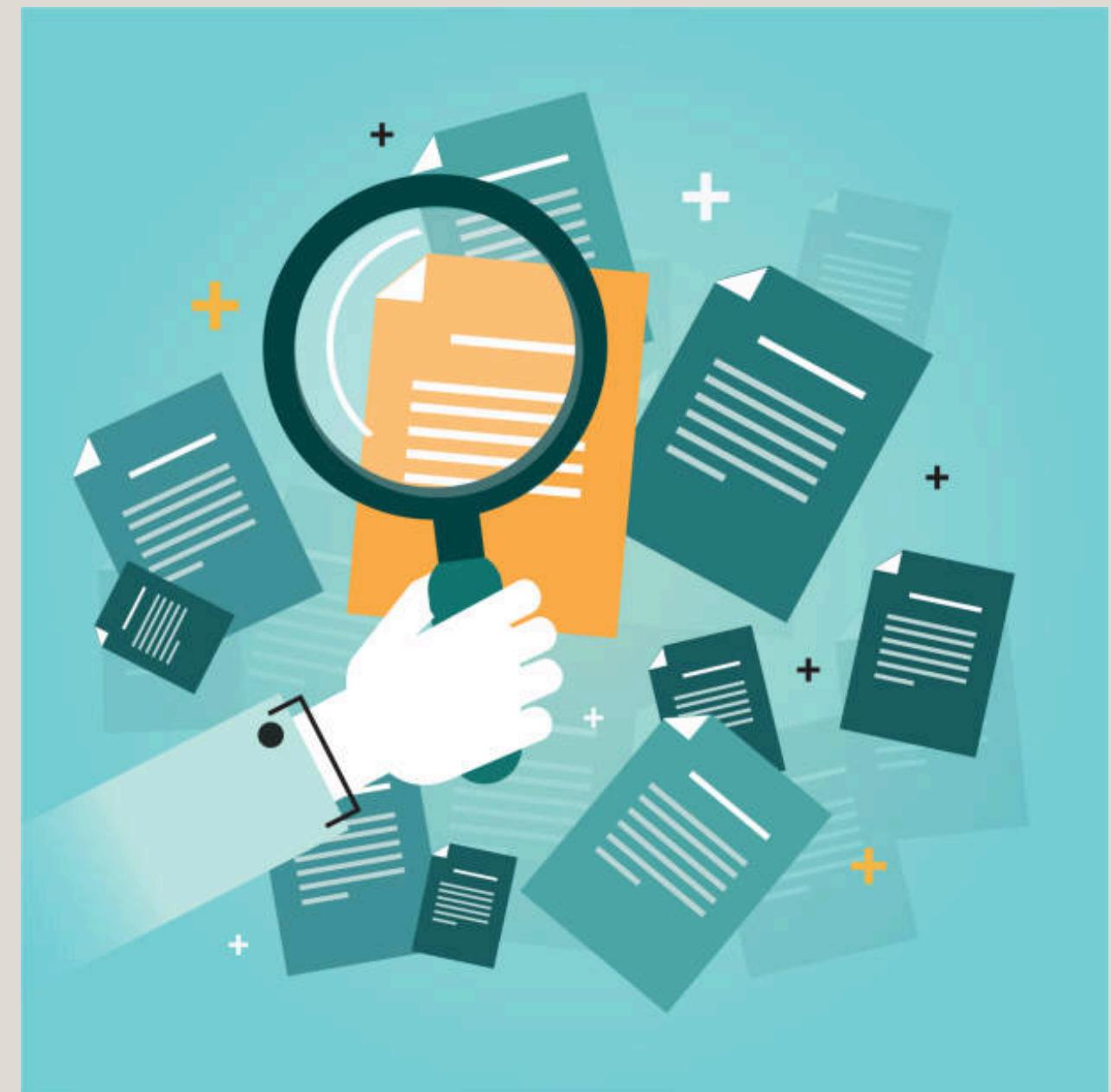


`is_canceled` (Target) tidak ada di Test Data

Karena `is_cancelled` tidak ada di `df_test`, maka kita akan menambahkan kolom `is_cancelled` berdasarkan isian dari `reservation_status`.

- Jika `reservation_status` = 'Check-Out', maka `is_cancelled` = 0
- Jika `reservation_status` != 'Check-Out', maka `is_cancelled` = 1

Exploratory Data Analysis



Statistical Summary



	count	mean	min	25%	50%	75%	max	std
is_canceled	118990.0	0.369493	0.0	0.0	0.0	1.0	1.0	0.48267
lead_time	118990.0	103.767064	0.0	18.0	69.0	160.0	737.0	106.679278
arrival_date_year	118990.0	2018.158602	2017.0	2018.0	2018.0	2019.0	2019.0	0.707192
arrival_date_week_number	118990.0	27.183713	1.0	16.0	28.0	38.0	53.0	13.609474
arrival_date_day_of_month	118990.0	15.786486	1.0	8.0	16.0	23.0	31.0	8.777821
stays_in_weekend_nights	118990.0	0.928969	0.0	0.0	1.0	2.0	19.0	0.999204
stays_in_week_nights	118990.0	2.501815	0.0	1.0	2.0	3.0	50.0	1.910087
adults	118990.0	1.855937	0.0	2.0	2.0	2.0	55.0	0.57999
children	118990.0	0.104009	0.0	0.0	0.0	0.0	10.0	0.398743
babies	118990.0	0.007942	0.0	0.0	0.0	0.0	10.0	0.097343
is_repeated_guest	118990.0	0.030465	0.0	0.0	0.0	0.0	1.0	0.171863
previous_cancellations	118990.0	0.085721	0.0	0.0	0.0	0.0	26.0	0.842812
previous_bookings_not_canceled	118990.0	0.137533	0.0	0.0	0.0	0.0	72.0	1.499924
booking_changes	118990.0	0.221615	0.0	0.0	0.0	0.0	21.0	0.653065
agent	118990.0	74.878586	0.0	7.0	9.0	152.0	535.0	107.168039
days_in_waiting_list	118990.0	2.328952	0.0	0.0	0.0	0.0	391.0	17.623754
adr	118990.0	101.893225	-6.38	69.8525	95.0	126.0	510.0	48.170937
required_car_parking_spaces	118990.0	0.062585	0.0	0.0	0.0	0.0	8.0	0.245422
total_of_special_requests	118990.0	0.572309	0.0	0.0	0.0	1.0	5.0	0.793219
reservation_status_date	118990	2018-07-31 13:44:31.424489216	2017-01-01 00:00:00	2018-02-02 00:00:00	2018-08-08 00:00:00	2019-02-09 00:00:00	2019-09-14 00:00:00	NaN
bookingID	118990.0	59495.31838	1.0	29748.25	59495.5	89241.75	118990.0	34349.41564
total_nights	118990.0	3.430784	0.0	2.0	3.0	4.0	69.0	2.559601

Rata-rata menginap:

- Weekend 0.93 malam
- Weekday 2.5 malam

Business recommendation untuk pertimbangan:

- Paket bundling 2 malam akhir pekan (diskon pada malam kedua, makan malam gratis, spa, dll)

Rata-rata repeated guest -> 3%

Business recommendation untuk pertimbangan:

- Buat loyalty program (point & rewards)
- Buat regular updates ke customer (newsletter terkait penawaran / promo khusus)

Rata-rata kebutuhan parkir-> 0.06 mobil

Business recommendation untuk pertimbangan:

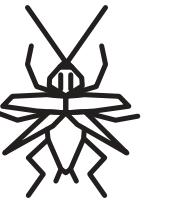
- Pemanfaatan ruang parkir menjadi fasilitas lain (taman, area rekreasi, dll)

Rata-rata special request -> 0.57 request

Business recommendation untuk pertimbangan:

- Kumpulkan & analisa request & feedback tamu, gunakan informasi untuk menyesuaikan penawaran & trend.

Statistical Summary



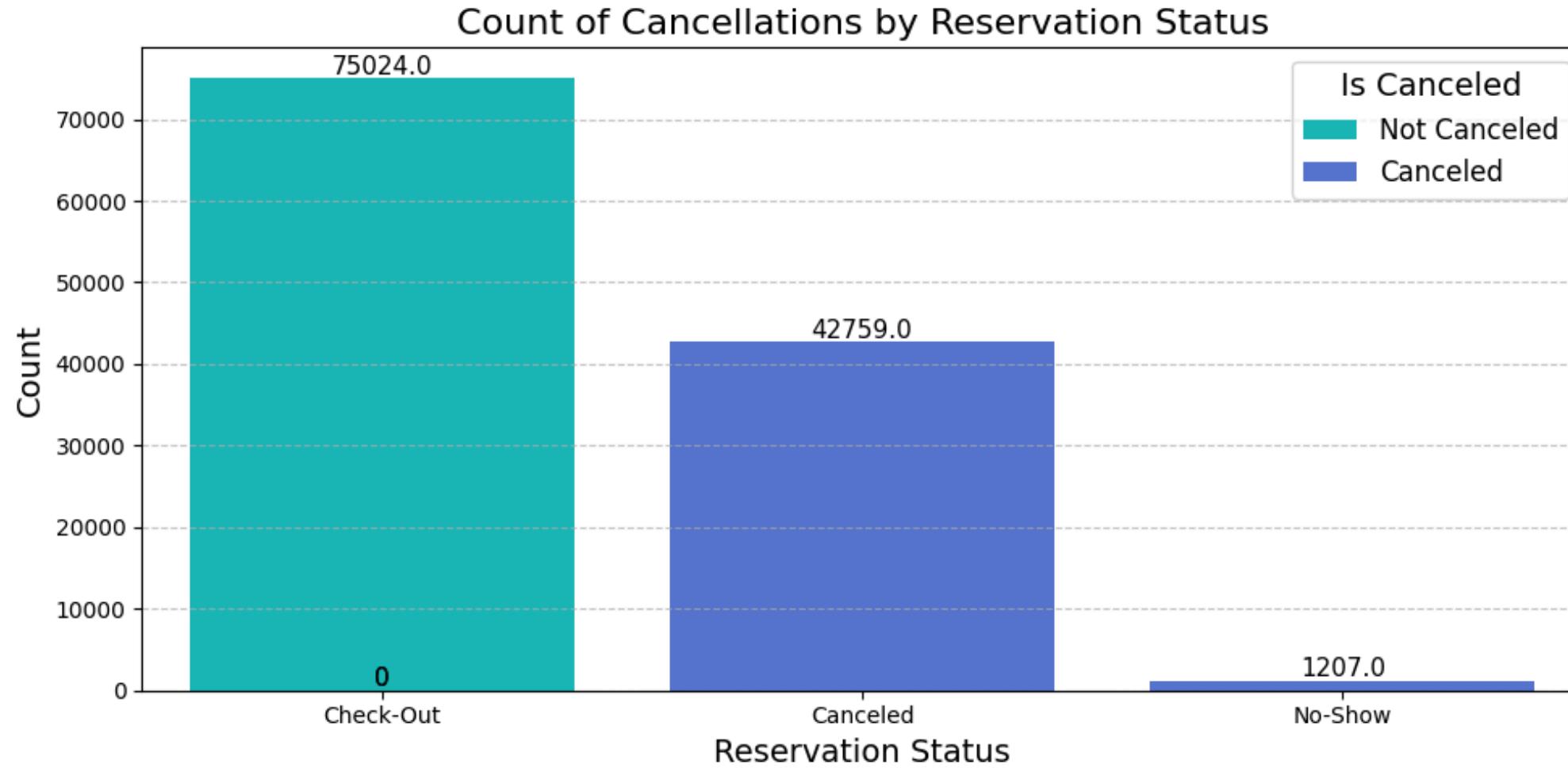
Terdapat 1 pesanan aneh dengan ADR negatif, total menginap 10 malam.

Bisa jadi merupakan kesalahan input

adr	118990.0	101.893225	-6.38	69.8525	95.0	126.0	510.0	48.170937
total_nights	10							

Decision -> Drop

EDA: Reservation Status



Observations:

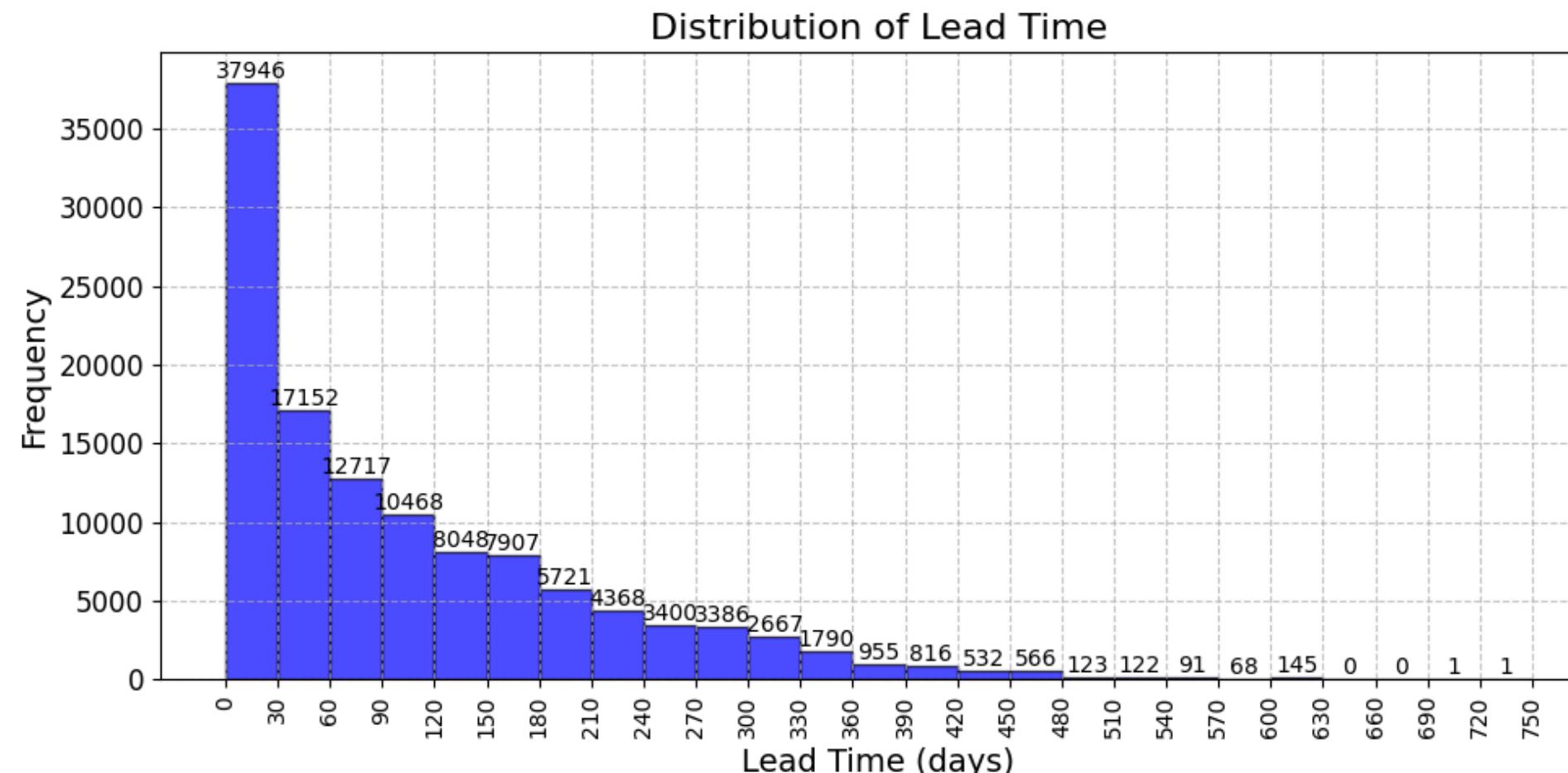
- Rata-rata dari tingkat pembatalan sekitar 36.9%, yang mana nilainya cukup besar
- 1207 pemesanan yang No-Show meskipun kecil namun merugikan (room bisa disewakan ke customer lain, jadi tidak bisa)

Business recommendations:

- Kebijakan pembatalan harus dievaluasi. Perusahaan bisa membuat atau menyesuaikan (jika sudah ada) biaya pembatalan untuk mengurangi pembatalan.
- Kebijakan seperti pembayaran DP atau biaya No-Show bisa dipertimbangkan agar mengurangi nilai No-Show

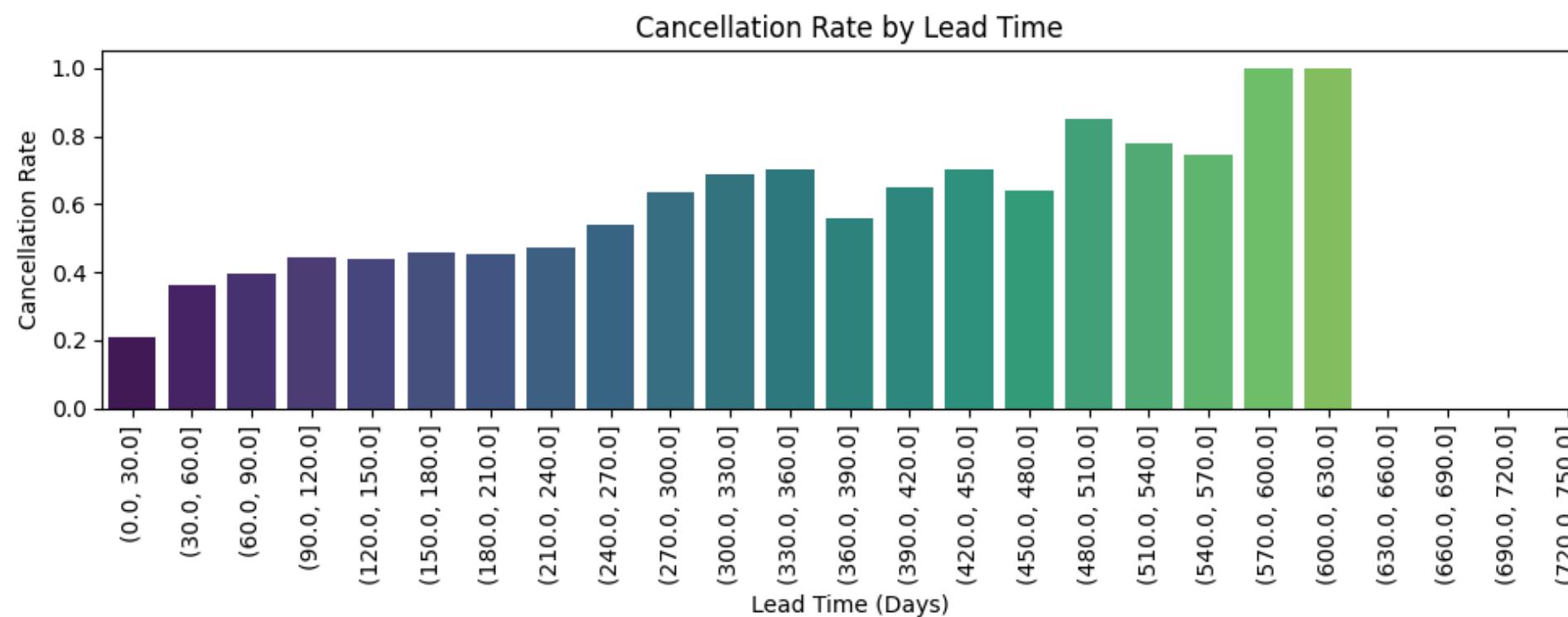
	count	mean	std	min	25%	50%	75%	max
is_canceled	118990.0	0.369493	0.48267	0.0	0.0	0.0	1.0	1.0

EDA: Lead Time



Observations:

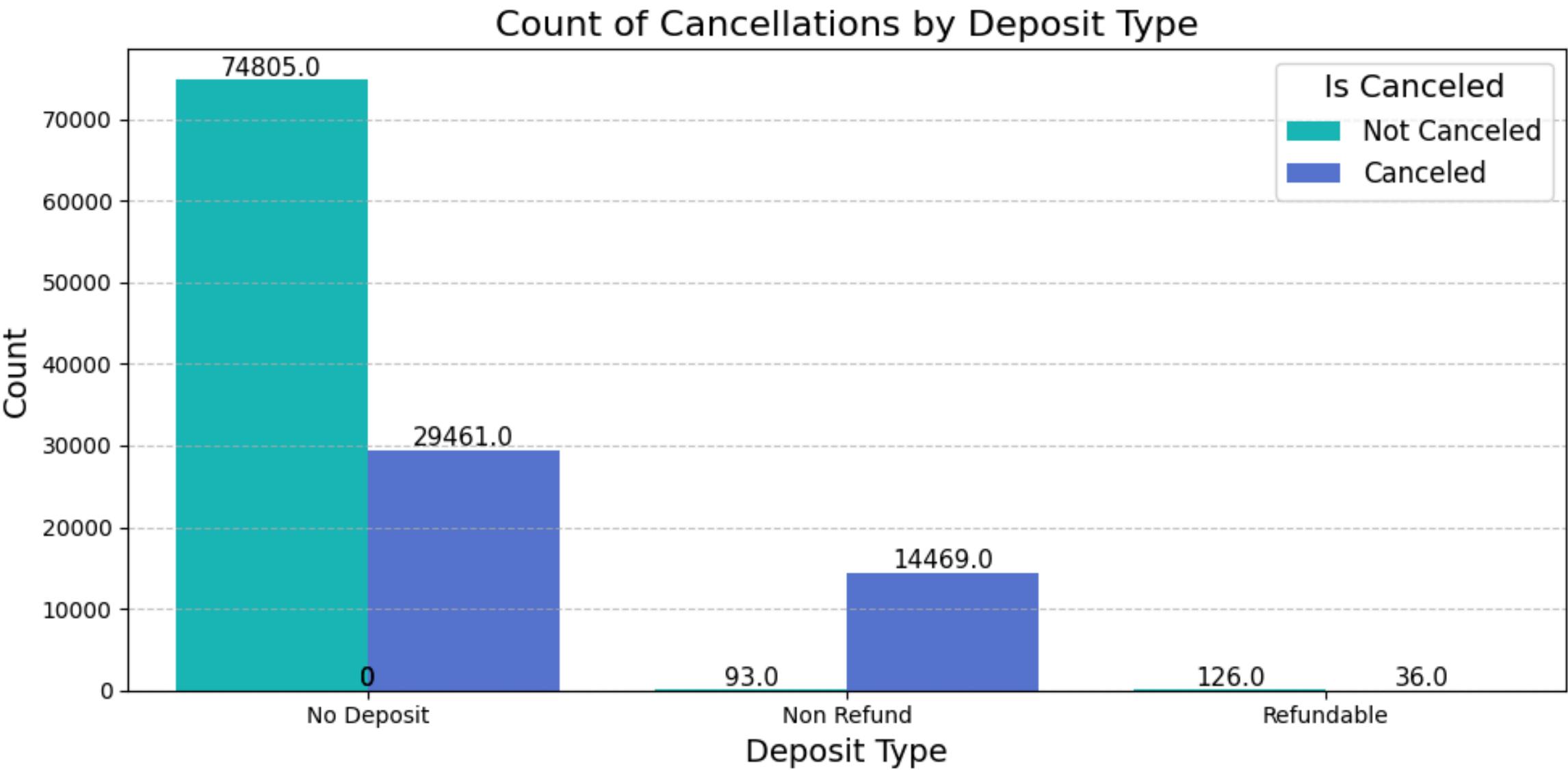
- Mayoritas lead_time yang pendek di bawah 30 hari (majoritas tamu memesan dalam waktu yang relatif dekat dengan tanggal kedatangan)
- Tingkat pembatalan (is_canceled) cenderung meningkat seiring dengan meningkatnya nilai lead_time
- lead_time di bawah 30 hari memiliki rata-rata tingkat pembatalan yang paling rendah
- lead_time yang sangat panjang (lebih dari 570 hari) memiliki tingkat pembatalan yang sangat tinggi



Business recommendations:

- Kebijakan pembatalan yang lebih ketat untuk pemesanan dengan lead_time yang panjang
- Loyalty program atau memberikan insentif untuk lead_time panjang dan tidak membatalkan

EDA: Deposit Type



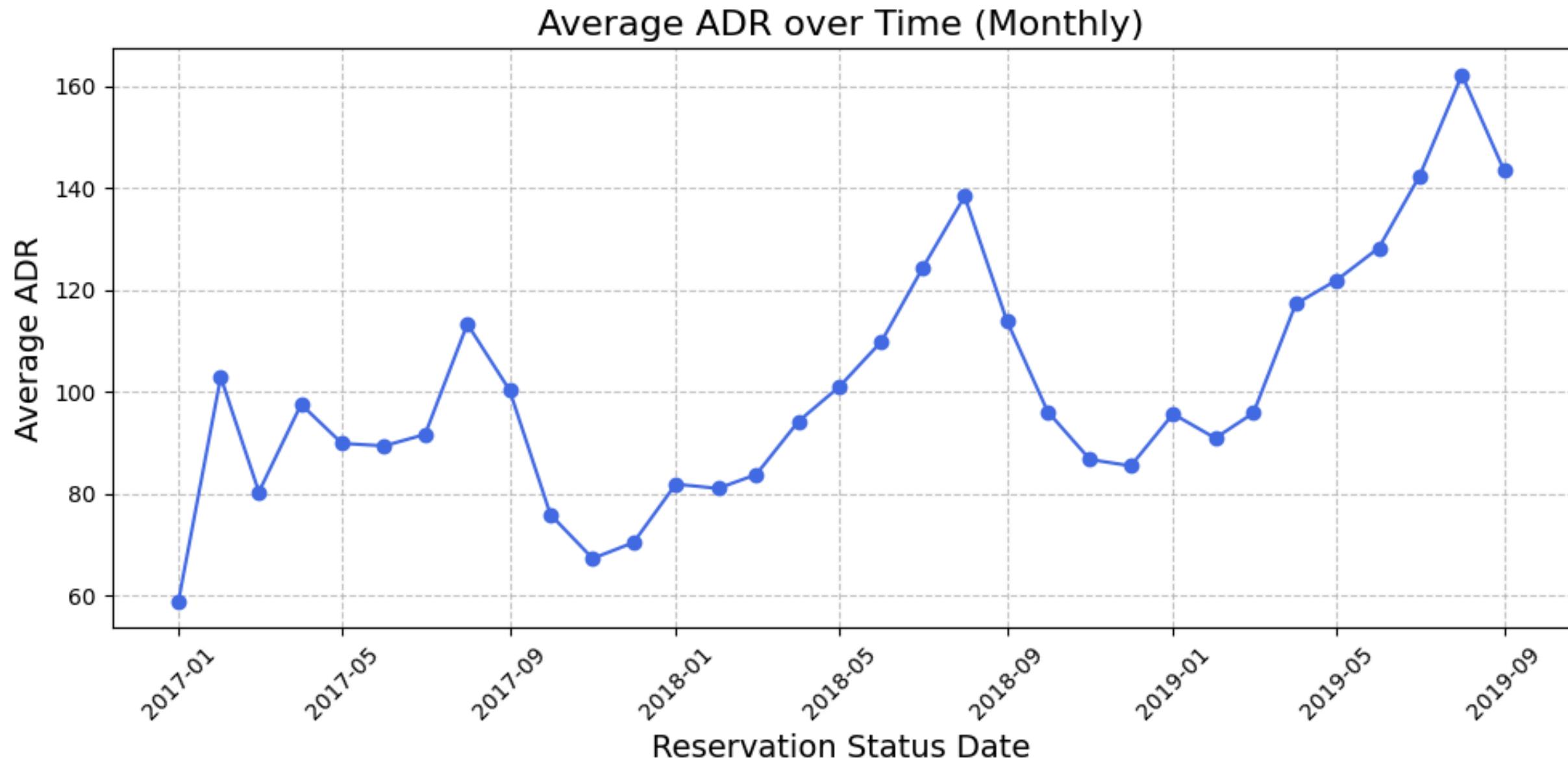
Observations:

- Mayoritas pemesanan dengan No Deposit tidak dibatalkan (Padahal fleksibilitasnya tinggi)
- Mayoritas pemesanan Non Refund dibatalkan (Artinya: meskipun deposit Non Refund dimaksudkan untuk mengurangi pembatalan, banyak tamu tetap membatalkan pemesanan)
- Mayoritas pemesanan dengan Refundable tidak dibatalkan (Sama seperti No Deposit, padahal fleksibilitasnya tinggi)

Business recommendations:

- Lakukan evaluasi ulang terhadap kebijakan Non Refund. Kebijakan deposit Non Refund ternyata tidak cukup efektif dalam mengurangi pembatalan. Bisa diganti dengan insentif khusus untuk tamu yang tidak membatalkan, dll.

EDA: ADR Over Time



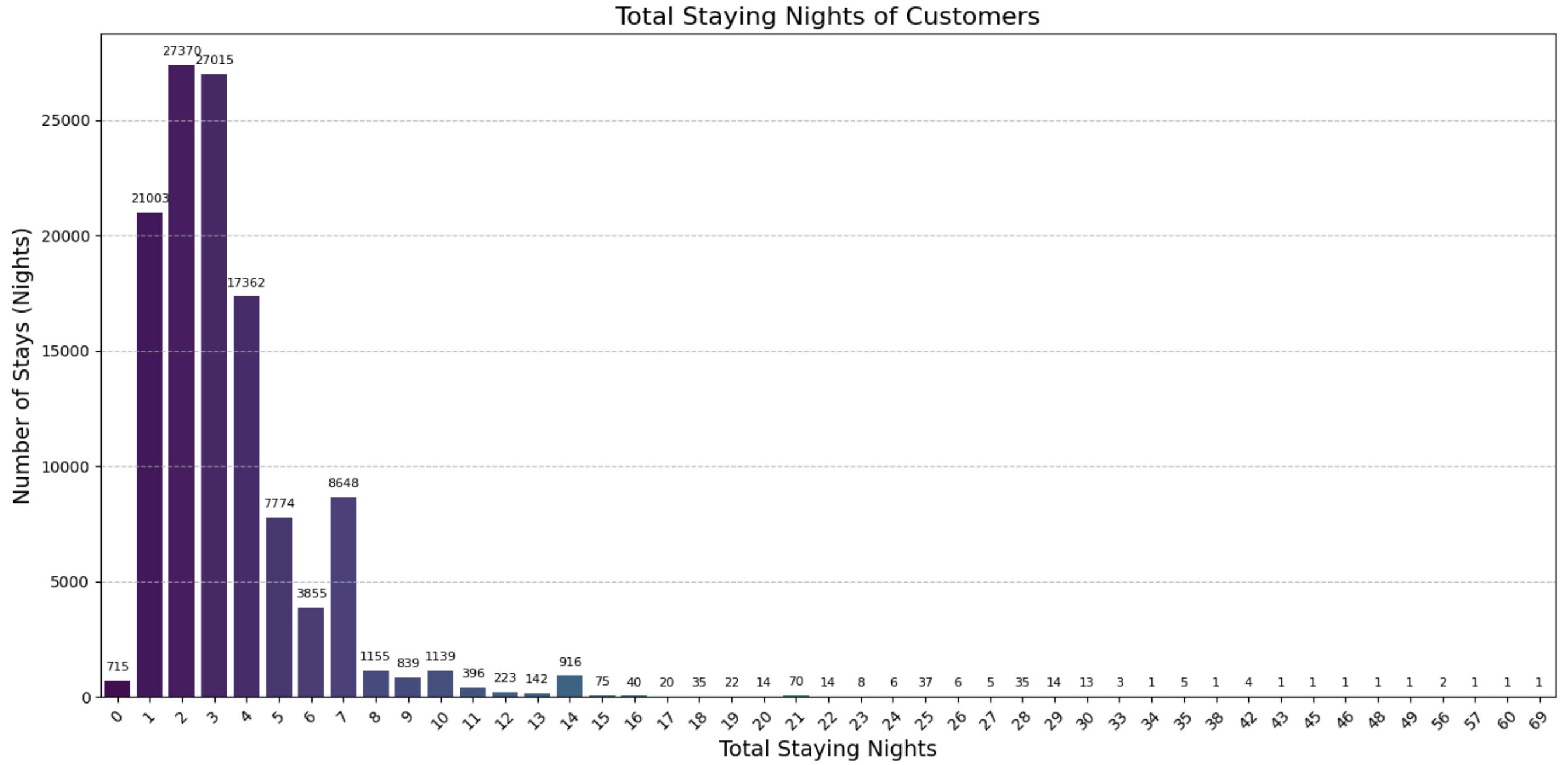
Observations:

- ADR tertinggi selalu pada bulan ke 8, dan tren positif dari bulan ke 3 menuju bulan ke 8.
- ADR selalu mengalami posisi menurun di bulan 8 ke atas, dan mulai meningkat lagi di bulan ke 3

Business recommendations:

- Promosi pada bulan-bulan ADR menurun. Misal: Promosi seperti diskon, paket bundling, atau penawaran khusus untuk acara.
- Melaksanakan strategi harga musiman (jika belum dilaksanakan). Misal: menaikkan harga selama bulan-bulan dengan permintaan tinggi dan turunkan harga selama bulan-bulan dengan permintaan rendah

EDA: Total Staying Nights



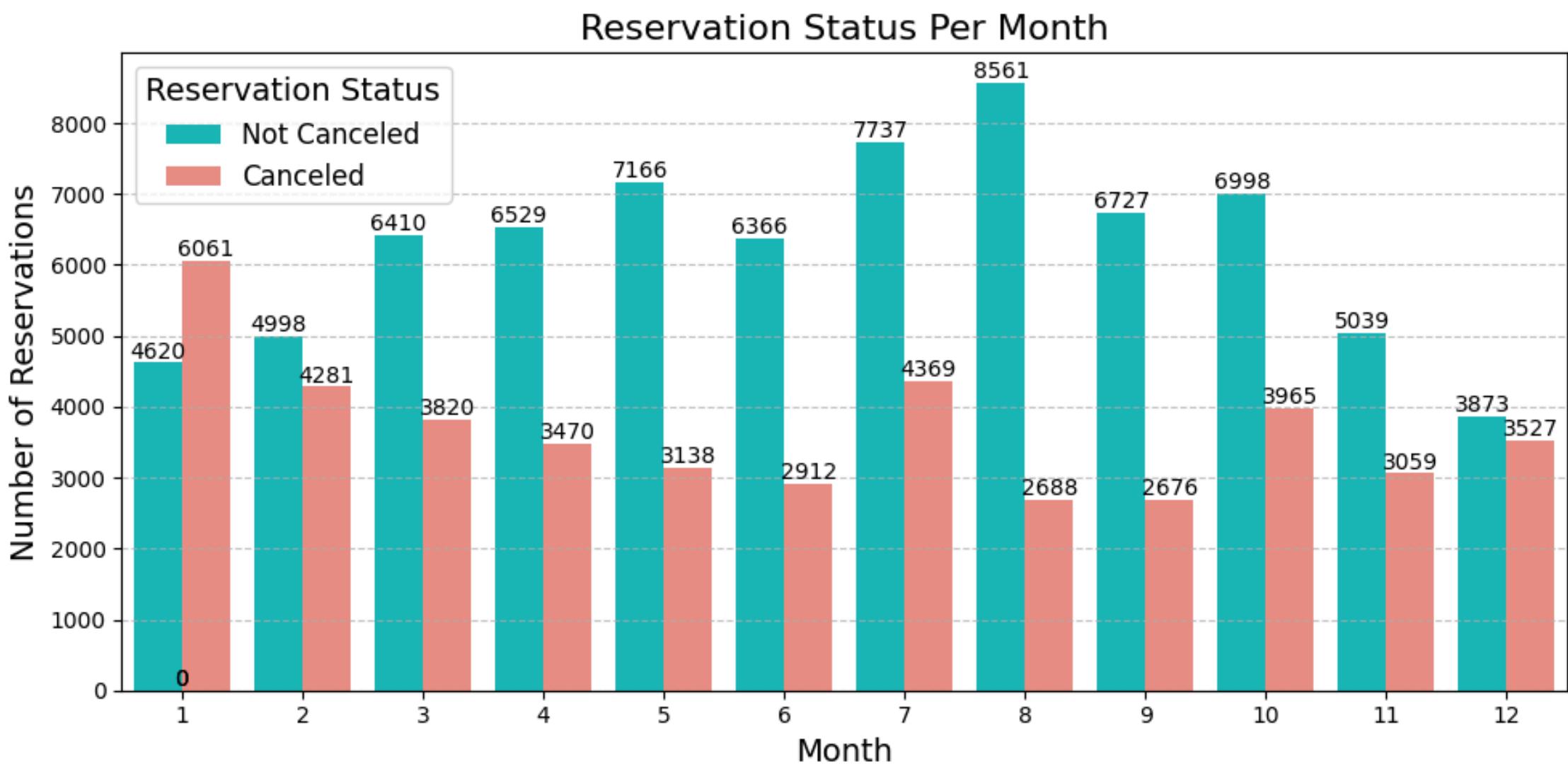
Observations:

- Mayoritas customer menginap 1 hingga 3 malam, dengan puncaknya pada 2 malam dan 3 malam menginap.
- Lama menginap lebih dari 7 malam sangat jarang terjadi
- Customer cenderung memilih menginap untuk waktu singkat (mungkin perjalanan bisnis)

Business recommendations:

- Diskon Progresif untuk menginap lama (>3 malam). Misal: diskon 10% untuk malam ke-4, 15% untuk malam ke-5, dan seterusnya
- Cek segmentasi pelanggan yang menginap lama & buat promosi yang ditargetkan

EDA: Reservation Status (Berdasarkan Bulan)



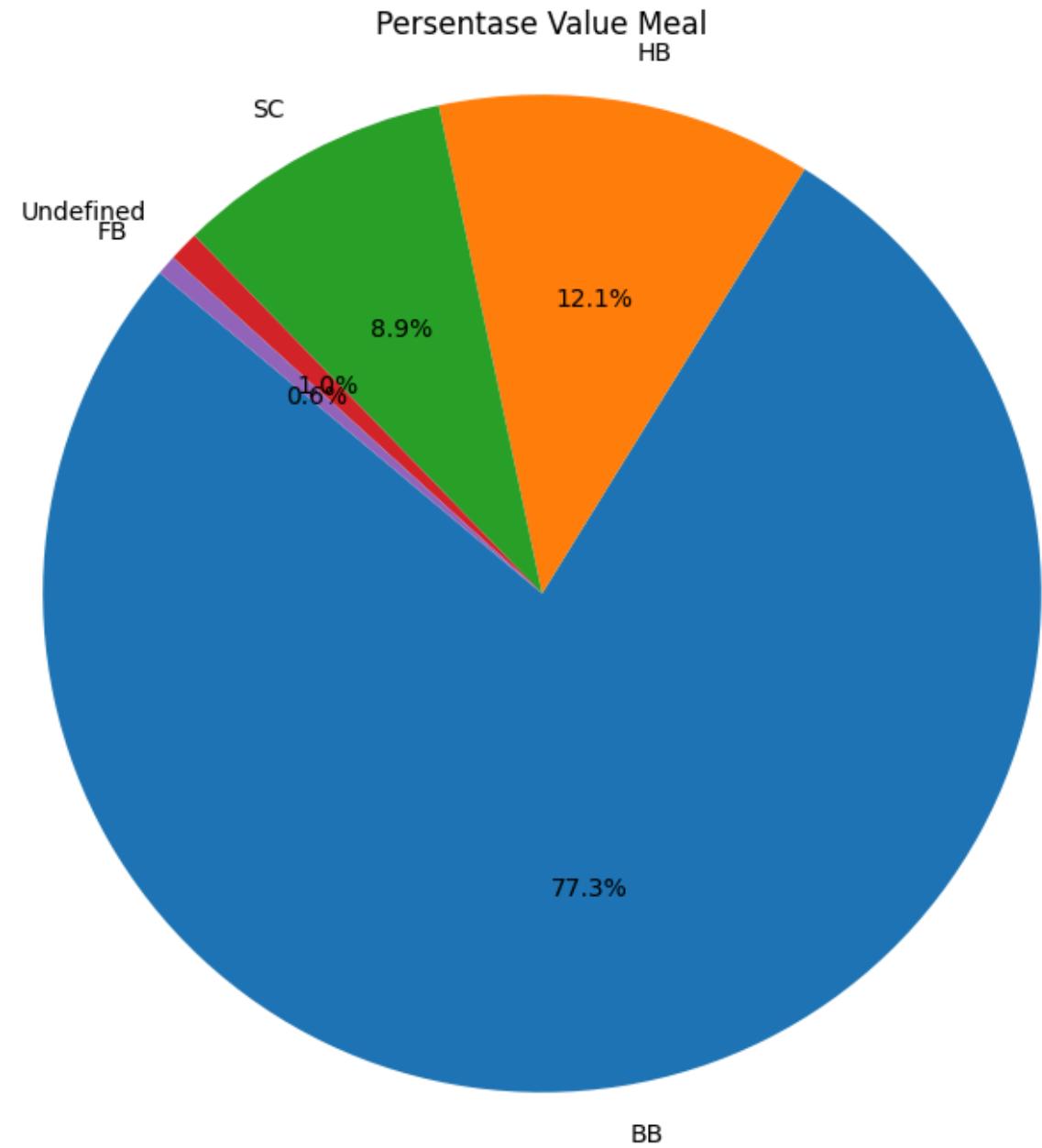
Observations:

- Bulan dengan rasio pembatalan (berdasarkan Not Canceled vs Canceled) tertinggi berdasarkan bulan pemesanannya adalah: 1 (Januari), 2 (Februari), 12 (Agustus)
- Bulan 7 (Juli) dan 8 (Agustus) memiliki jumlah pemesanan Not Canceled paling tinggi. Sepertinya ini ada hubungan dengan liburan musim panas

Business recommendations:

- Promosi pada bulan dengan tingkat pembatalan tinggi (Januari, Februari, Agustus). Misal: Incentif khusus untuk tamu yang tidak membatalkan
- Promosi khusus untuk meningkatkan pendapatan pada bulan Not Canceled paling tinggi (Mei, Juli, Agustus). Misal: Paket bundle wisata, liburan, dll untuk menaikkan pendapatan.

EDA: Persentase Meals



Observations:

- Bed and Breakfast (BB) merupakan mayoritas dengan persentase 77.3%. Hal ini menunjukkan bahwa kebanyakan tamu lebih suka mendapatkan sarapan / sarapan di kamar.

Business recommendations:

- Mayoritas BB -> Memastikan kalau kualitas sarapan pagi mereka baik, agar customer puas & meningkatkan kunjungan tamu berulang.

Multicollinearity Check



	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies	is_repeated_guest	previous_cancellations	previous_bookings_notCanceled	booking_changes	agent	days_in_waiting_list	adr	required_car_parking_spaces	total_of_special_requests	bookingID	total_nights
is_canceled	-1.00	0.29	0.02	0.01	-0.00	0.00	0.02	0.06	0.01	-0.03	-0.10	0.11	-0.06	-0.14	-0.05	0.05	0.05	-0.20	-0.23	-0.00	0.02
lead_time	0.29	1.00	0.04	0.13	0.00	0.09	0.17	0.12	-0.04	-0.02	-0.14	0.08	-0.07	0.00	-0.01	0.17	-0.06	-0.12	-0.09	0.00	0.16
arrival_date_year	-0.02	0.04	1.00	-0.54	0.00	0.02	0.03	0.03	0.05	-0.01	0.02	-0.12	0.03	0.03	0.06	-0.06	0.21	-0.01	0.11	0.00	0.03
arrival_date_week_number	0.01	0.13	-0.54	1.00	0.07	0.02	0.02	0.03	0.01	0.01	-0.03	0.03	-0.02	0.01	-0.02	0.02	0.08	0.00	0.03	0.00	0.02
arrival_date_day_of_month	-0.00	0.00	0.00	0.07	1.00	-0.02	-0.03	-0.00	0.01	-0.00	-0.01	-0.03	-0.00	0.01	-0.00	0.02	0.03	0.01	0.00	0.00	-0.03
stays_in_weekend_nights	-0.00	0.09	0.02	0.02	-0.02	1.00	0.50	0.09	0.05	0.02	-0.08	-0.01	-0.04	0.06	0.16	-0.05	0.05	-0.02	0.07	-0.00	0.76
stays_in_week_nights	-0.02	0.17	0.03	0.02	-0.03	0.50	1.00	0.09	0.04	0.02	-0.10	-0.01	-0.05	0.10	0.20	-0.00	0.07	-0.03	0.07	0.00	0.94
adults	-0.06	0.12	0.03	0.03	-0.00	0.09	0.09	1.00	0.03	0.02	-0.15	-0.01	-0.11	-0.05	0.03	-0.01	0.24	0.01	0.12	0.00	0.11
children	-0.01	-0.04	0.05	0.01	0.05	0.04	0.03	1.00	0.02	-0.03	-0.02	-0.02	0.05	0.05	-0.03	0.34	0.06	0.08	0.00	0.05	
babies	-0.03	-0.02	-0.01	0.01	-0.00	0.02	0.02	0.02	1.00	-0.01	-0.01	-0.01	0.08	0.03	-0.01	0.03	0.04	0.10	-0.00	0.02	
is_repeated_guest	-0.10	-0.14	0.02	-0.03	-0.01	-0.08	-0.10	-0.15	-0.03	-0.01	1.00	0.07	0.43	0.02	-0.05	-0.02	-0.14	0.08	0.02	-0.00	-0.10
previous_cancellations	0.11	0.08	-0.12	0.03	-0.03	-0.01	-0.01	-0.01	-0.02	-0.01	0.07	1.00	0.15	-0.03	-0.02	0.01	-0.07	-0.02	-0.05	0.00	-0.01
previous_bookings_notCanceled	-0.06	-0.07	0.03	-0.02	-0.00	-0.04	-0.05	-0.11	-0.02	-0.01	0.43	0.15	1.00	0.01	-0.05	-0.01	-0.08	0.05	0.04	-0.00	-0.05
booking_changes	-0.14	0.00	0.03	0.01	0.01	0.06	0.10	-0.05	0.05	0.08	0.02	-0.03	0.01	1.00	0.04	-0.01	0.02	0.07	0.05	-0.00	0.10
agent	-0.05	-0.01	0.06	-0.02	-0.00	0.16	0.20	0.03	0.05	0.03	-0.05	-0.02	-0.05	0.04	1.00	-0.04	0.02	0.12	0.06	0.00	0.21
days_in_waiting_list	-0.05	0.17	-0.06	0.02	0.02	-0.05	-0.00	-0.01	-0.03	-0.01	-0.02	0.01	-0.01	-0.01	-0.04	1.00	-0.04	-0.03	-0.08	-0.00	-0.02
adr	-0.05	-0.06	0.21	0.08	0.03	0.05	0.07	0.24	0.34	0.03	-0.14	-0.07	-0.08	0.02	0.02	-0.04	1.00	0.06	0.18	0.00	0.07
required_car_parking_spaces	-0.20	-0.12	-0.01	0.00	0.01	-0.02	-0.03	0.01	0.06	0.04	0.08	-0.02	0.05	0.07	0.12	-0.03	0.06	1.00	0.08	0.00	-0.03
total_of_special_requests	-0.23	-0.09	0.11	0.03	0.00	0.07	0.07	0.12	0.08	0.10	0.02	-0.05	0.04	0.05	0.06	-0.08	0.18	0.08	1.00	0.00	0.08
bookingID	-0.00	-0.00	0.00	0.00	-0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	0.00	1.00	-0.00	
total_nights	-0.02	0.16	0.03	0.02	-0.03	0.76	0.94	0.11	0.05	0.02	-0.10	-0.01	-0.05	0.10	0.21	-0.02	0.07	-0.03	0.08	-0.00	1.00
	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies	is_repeated_guest	previous_cancellations	previous_bookings_notCanceled	booking_changes	agent	days_in_waiting_list	adr	required_car_parking_spaces	total_of_special_requests	bookingID	total_nights

Observasi korelasi:

- `total_nights` & `stays_in_weekend_nights` = 0.76
- `total_nights` & `stays_in_week_nights` = 0.94

Keputusan:

- `total_nights` adalah kombinasi dari `stays_in_weekend_nights` & `stays_in_week_nights`, kita bisa keep `total_nights` dan membuang kedua kolom asli.

feature importance antara `total_nights`, `stays_in_weekend_nights` dan `stays_in_week_nights`

Feature	Importance
total_nights	0.575694
stays_in_week_nights	0.298061
stays_in_weekend_nights	0.126245

Keputusan:

- Keep `total_nights`
- Drop `stays_in_weekend_nights` & `stays_in_week_nights`,

Feature Encoding



Observations:

- Seluruh kolom categorical >2 unique values dan tidak ordinal

```
hotel           [Crystal Cove, Barbados Barbados, Greensboro C...
arrival_date_month  [January, December, May, October, September, M...
meal             [BB, HB, SC, FB, Undefined]
country          [ITA, PRT, BEL, DNK, GBR, IRL, ESP, DEU, RUS, ...
market_segment   [Online TA, Groups, Corporate, Offline TA/T0, ...
distribution_channel [TA/T0, Corporate, Direct, GDS, Undefined]
reserved_room_type [A, D, E, G, F, H, B, C, P, L]
assigned_room_type [A, D, E, C, F, G, H, B, K, I, P]
deposit_type     [No Deposit, Non Refund, Refundable]
customer_type    [Transient, Contract, Transient-Party, Group]
dtype: object
```

```
hotel: 64
arrival_date_month: 12
country: 165
market_segment: 8
```

Decision:

- Frequency encoding berdasarkan persentase kemunculan value pada tiap kolom.

Drop Features



bookingID

ID Booking tidak akan berguna untuk prediction modeling karena tidak memberikan pola apapun.

assigned_room_type

Assigned Room Type tidak membantu proses prediction. Karena tipe ruangan yang diberikan (final) kondisinya adalah ketika pelanggan sudah sampai di hotel. Keputusan untuk membatalkan biasanya dibuat sebelum kamar diberikan.

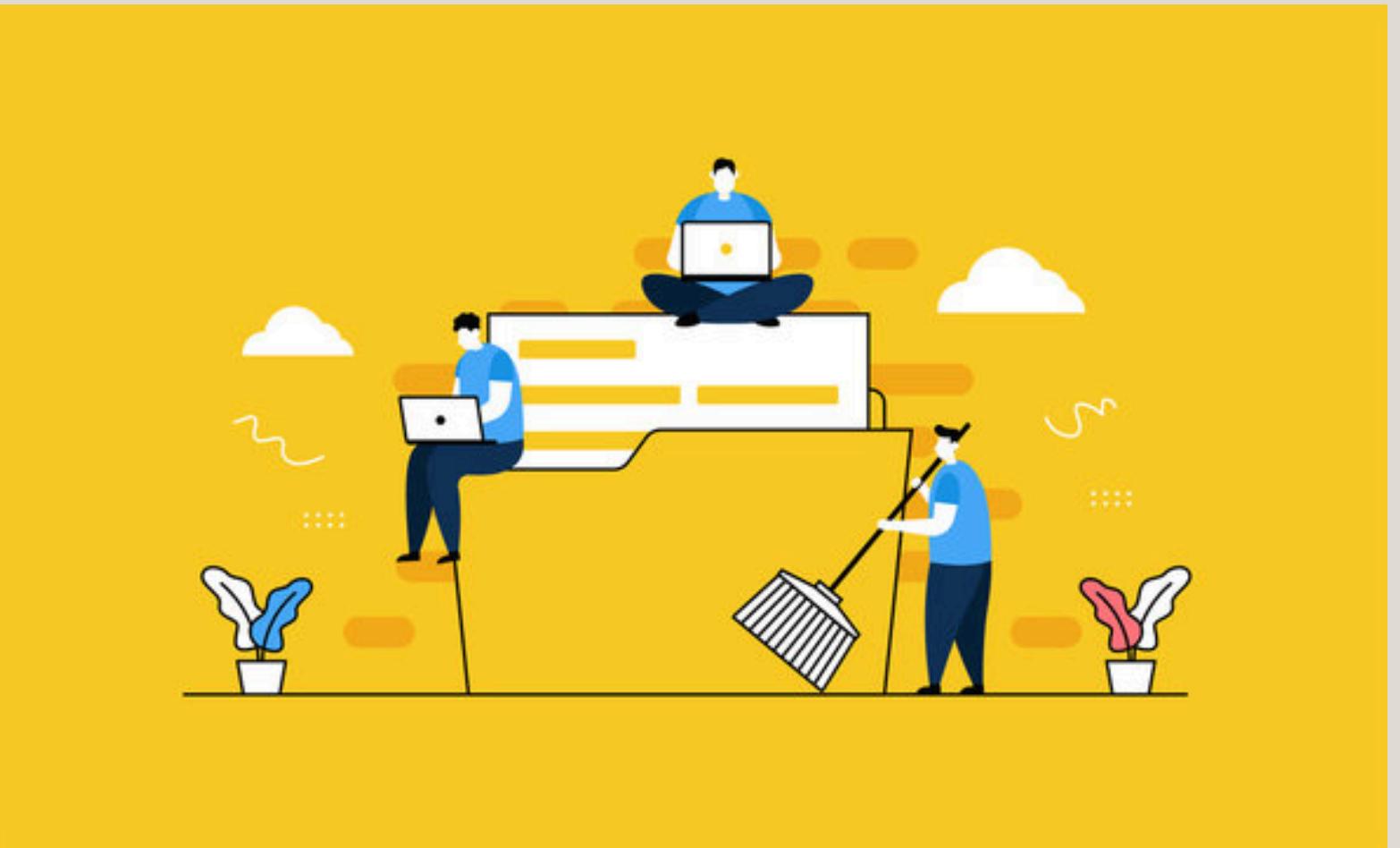
reservation_status_date

reservation_status_date adalah tanggal perubahan status terakhir dari reservasi. Ini dapat berupa tanggal pembatalan, tanggal tidak datang (no show), atau tanggal check-out.

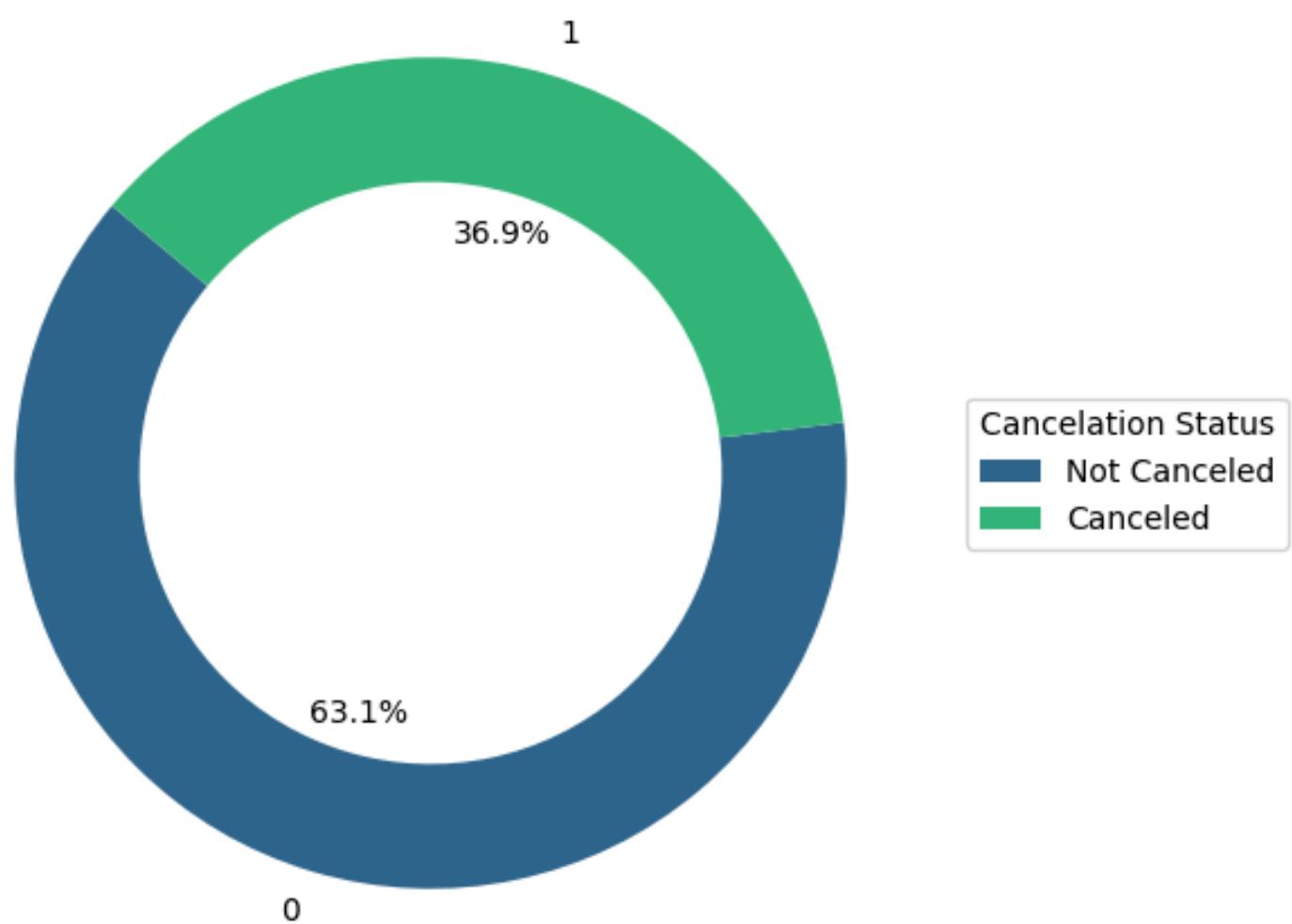
Jika digunakan dalam prediksi, ini bisa menyebabkan data leakage.

- Jika $\text{reservation_status_date} < \text{arrival_date}$ = Cancel
- Jika $\text{reservation_status_date} = \text{arrival_date}$ = No Show
- Jika $\text{reservation_status_date} > \text{arrival_date}$ = Check Out

Modeling



Distribusi Cancelation Status



Karena target memiliki ketidakseimbangan kelas antara Not Canceled dan Canceled, maka metric evaluasi utama yang digunakan di sini adalah F1-Score, dibanding accuracy.

Modeling Result (Train Data)

Results on Train Data

Model yang dicoba:

- LightGBM
- Random Forest
- Logistic Regression

LightGBM							Random Forest				
LightGBM Train F1: 0.8296474663625035					RandomForest Train F1: 0.9991459943066288						
LightGBM Train Classification Report:					RandomForest Train Classification Report:						
	precision	recall	f1-score	support		precision	recall	f1-score	support		
0	0.89	0.93	0.91	42044		0	1.00	1.00	1.00	42044	
1	0.86	0.80	0.83	24590		1	1.00	1.00	1.00	24590	
accuracy			0.88	66634	accuracy				1.00	66634	
macro avg	0.88	0.86	0.87	66634	macro avg	1.00	1.00	1.00	66634		
weighted avg	0.88	0.88	0.88	66634	weighted avg	1.00	1.00	1.00	66634		

Logistic Regression

Logistic Regression					SVM						
LogisticRegression Train F1: 0.6921265296698222					SVM Train F1: 0.7904209884075657						
LogisticRegression Train Classification Report:					SVM Train Classification Report:						
	precision	recall	f1-score	support		precision	recall	f1-score	support		
0	0.80	0.91	0.85	42044		0	0.86	0.92	0.89	42044	
1	0.80	0.61	0.69	24590		1	0.85	0.74	0.79	24590	
accuracy			0.80	66634	accuracy				0.86	66634	
macro avg	0.80	0.76	0.77	66634	macro avg	0.85	0.83	0.84	66634		
weighted avg	0.80	0.80	0.79	66634	weighted avg	0.86	0.86	0.85	66634		

Modeling Result (Test Data)

Results on Test Data

LightGBM

LightGBM Test F1: 0.789213078528461

LightGBM Test Classification Report:

	precision	recall	f1-score	support
0	0.85	0.93	0.89	22508
1	0.86	0.73	0.79	13189
accuracy			0.86	35697
macro avg	0.86	0.83	0.84	35697
weighted avg	0.86	0.86	0.85	35697

Random Forest

RandomForest Test F1: 0.8194821508164837

RandomForest Test Classification Report:

	precision	recall	f1-score	support
0	0.87	0.94	0.90	22508
1	0.88	0.77	0.82	13189
accuracy			0.88	35697
macro avg	0.88	0.85	0.86	35697
weighted avg	0.88	0.88	0.87	35697

Logistic Regression

LogisticRegression Test F1: 0.7022400136274595

LogisticRegression Test Classification Report:

	precision	recall	f1-score	support
0	0.81	0.91	0.85	22508
1	0.80	0.63	0.70	13189
accuracy			0.80	35697
macro avg	0.80	0.77	0.78	35697
weighted avg	0.80	0.80	0.80	35697

SVM

SVM Test F1: 0.7808475957620212

SVM Test Classification Report:

	precision	recall	f1-score	support
0	0.85	0.92	0.89	22508
1	0.84	0.73	0.78	13189
accuracy			0.85	35697
macro avg	0.85	0.82	0.83	35697
weighted avg	0.85	0.85	0.85	35697

Metric terbaik pada test data ada pada
model Random Forest

Overfit Caution

Results on Random Forest

Train Data

RandomForest Train F1: 0.9991459943066288				
RandomForest Train Classification Report:				
	precision	recall	f1-score	support
0	1.00	1.00	1.00	42044
1	1.00	1.00	1.00	24590
accuracy			1.00	66634
macro avg	1.00	1.00	1.00	66634
weighted avg	1.00	1.00	1.00	66634

Test Data

RandomForest Test F1: 0.8194821508164837				
RandomForest Test Classification Report:				
	precision	recall	f1-score	support
0	0.87	0.94	0.90	22508
1	0.88	0.77	0.82	13189
accuracy			0.88	35697
macro avg	0.88	0.85	0.86	35697
weighted avg	0.88	0.88	0.87	35697

Overfit

Random Forest menunjukkan overfitting yang sangat jelas. F1 score pada data training mendekati 1 (0.9991), tetapi turun signifikan pada data test (0.8195).

Hyperparameter Tuning (Randomized)

Model Default (Train Data)

```
RandomForest Train F1: 0.9991459943066288
RandomForest Train Classification Report:
precision    recall    f1-score   support
0            1.00     1.00      1.00     42044
1            1.00     1.00      1.00     24590
accuracy          1.00      1.00      1.00     66634
macro avg       1.00     1.00      1.00     66634
weighted avg    1.00     1.00      1.00     66634
```

Model Default (Test Data)

```
RandomForest Test F1: 0.8194821508164837
RandomForest Test Classification Report:
precision    recall    f1-score   support
0            0.87     0.94      0.90     22508
1            0.88     0.77      0.82     13189
accuracy          0.88      0.85      0.88     35697
macro avg       0.88     0.85      0.86     35697
weighted avg    0.88     0.88      0.87     35697
```

Model Tuned (Train Data)

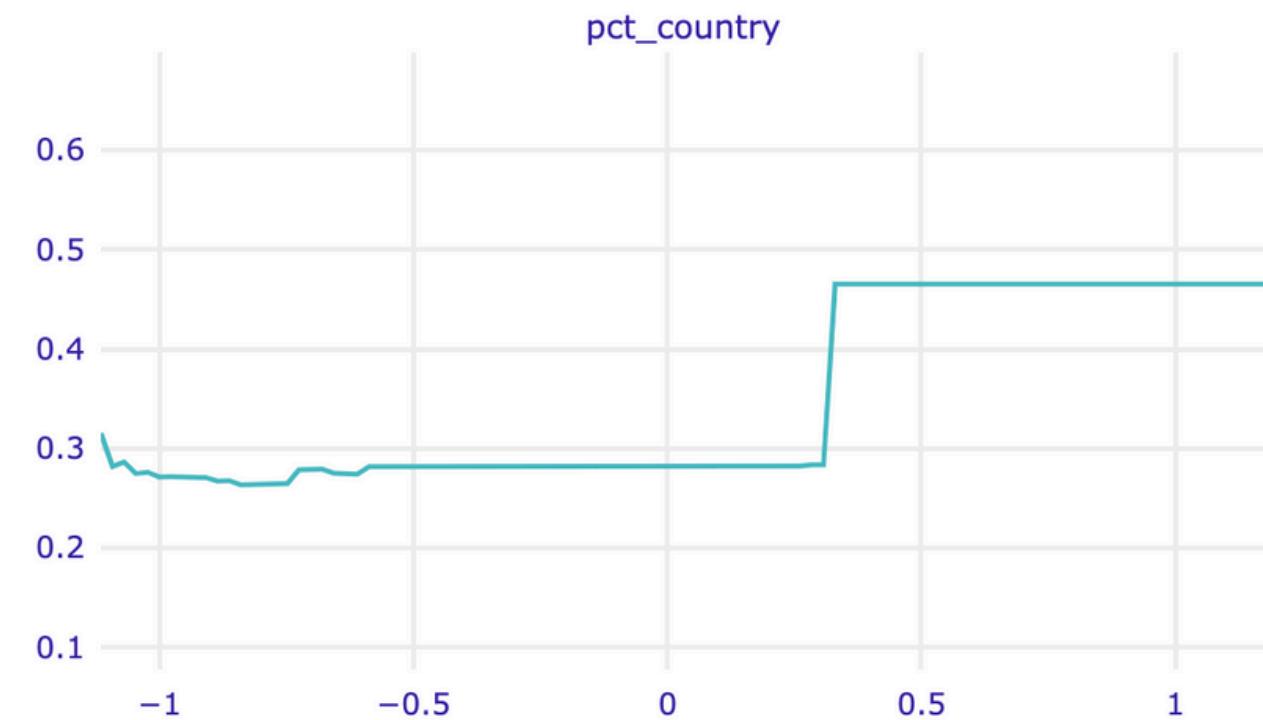
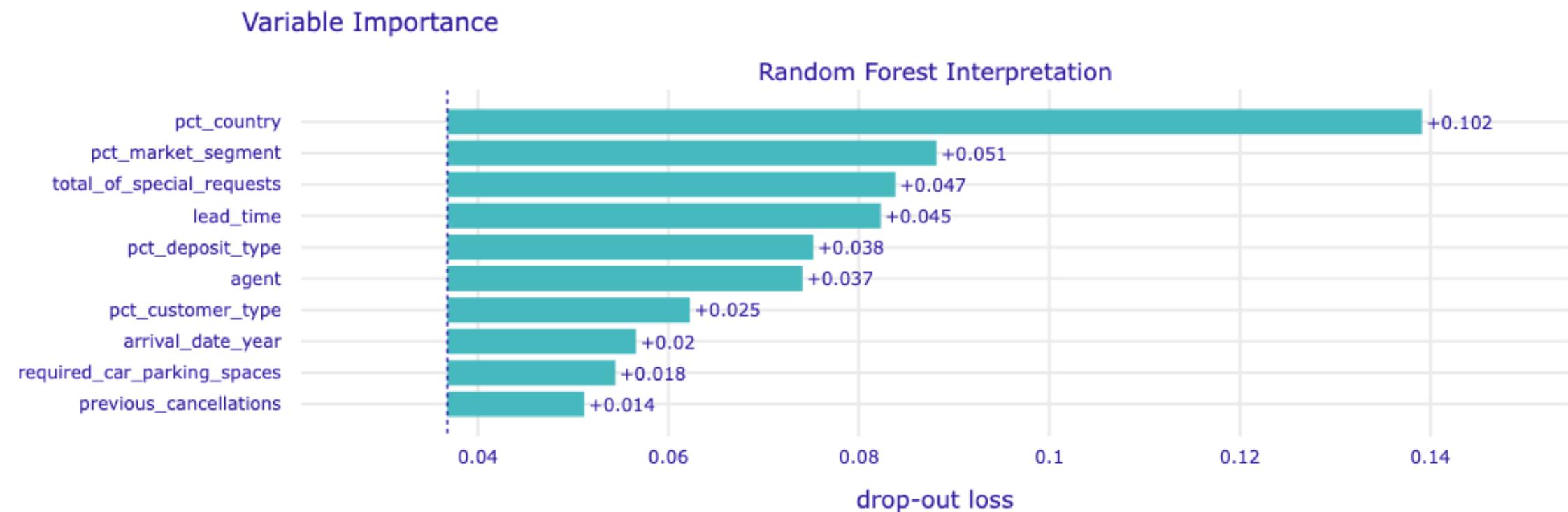
```
Train F1: 0.8287023962112489
Classification Report:
precision    recall    f1-score   support
0            0.88     0.94      0.91     42044
1            0.89     0.78      0.83     24590
accuracy          0.88      0.88      0.88     66634
macro avg       0.88     0.86      0.87     66634
weighted avg    0.88     0.88      0.88     66634
```

Model Tuned (Test Data)

```
Random Forest Test F1: 0.7948345305817381
Random Forest Test Classification Report:
precision    recall    f1-score   support
0            0.85     0.94      0.90     22508
1            0.88     0.73      0.79     13189
accuracy          0.86      0.86      0.86     35697
macro avg       0.87     0.83      0.85     35697
weighted avg    0.86     0.86      0.86     35697
```

- Randomized search berhasil menemukan parameter yang lebih baik pada training model.
- Setelah hyperparameter tuning, model menjadi lebih realistik dan tidak overfit
- Meskipun F1 score pada data test setelah hyperparameter tuning sedikit lebih rendah daripada model default, model yang telah di-tuning menunjukkan performa yang lebih seimbang dan lebih dapat diandalkan.

Feature Importance (Country)

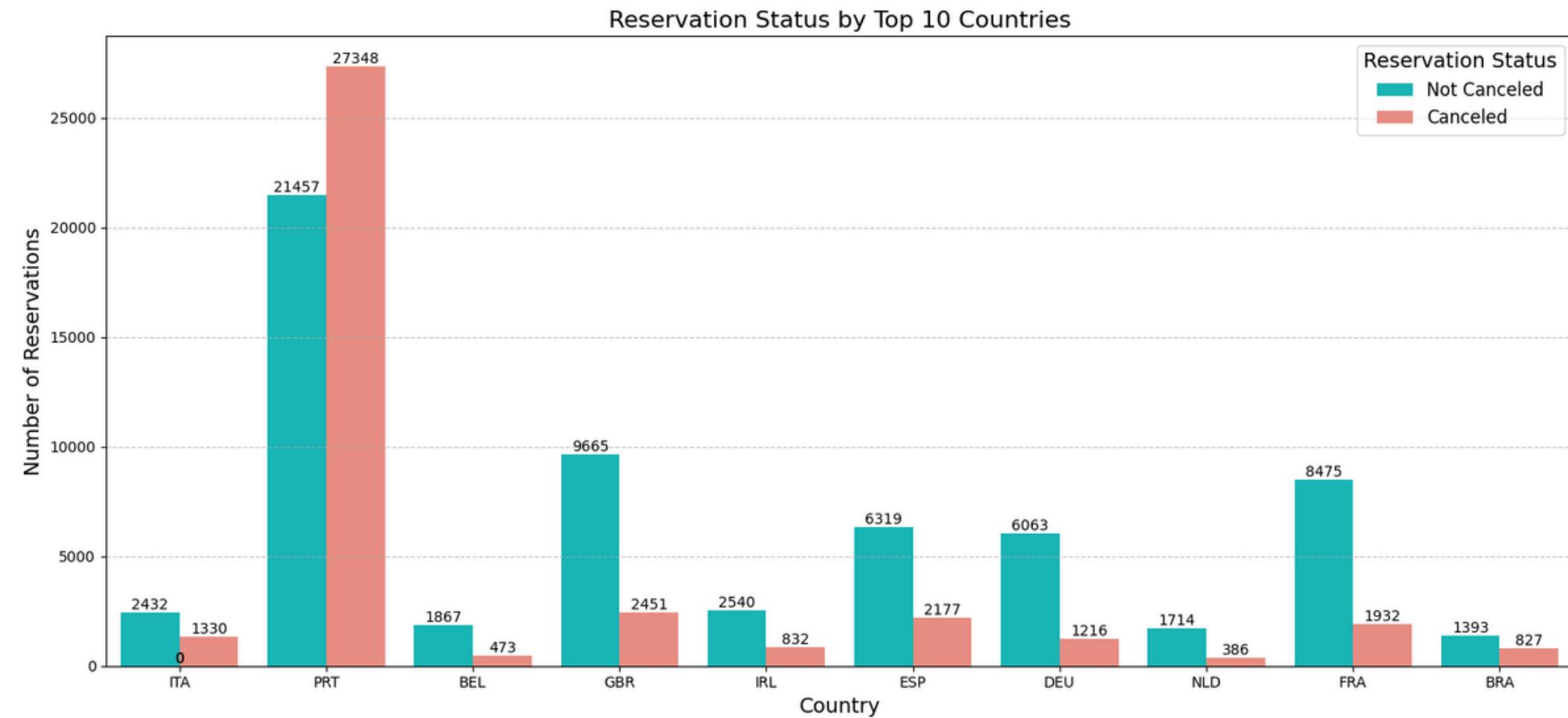


Insight

- `pct_country` menempati posisi paling atas bagi feature importance model.
- Menunjukkan pengaruh besar dari persentase negara asal tamu terhadap kemungkinan pembatalan reservasi.
- Tamu dari negara tertentu mungkin memiliki kecenderungan yang lebih tinggi untuk membatalkan reservasi.

Selanjutnya akan dilakukan analisis lebih lanjut terhadap hubungan antara negara dengan kesuksesan pemesanan (dibatalkan atau tidaknya)

Feature Importance (Cont. Country)



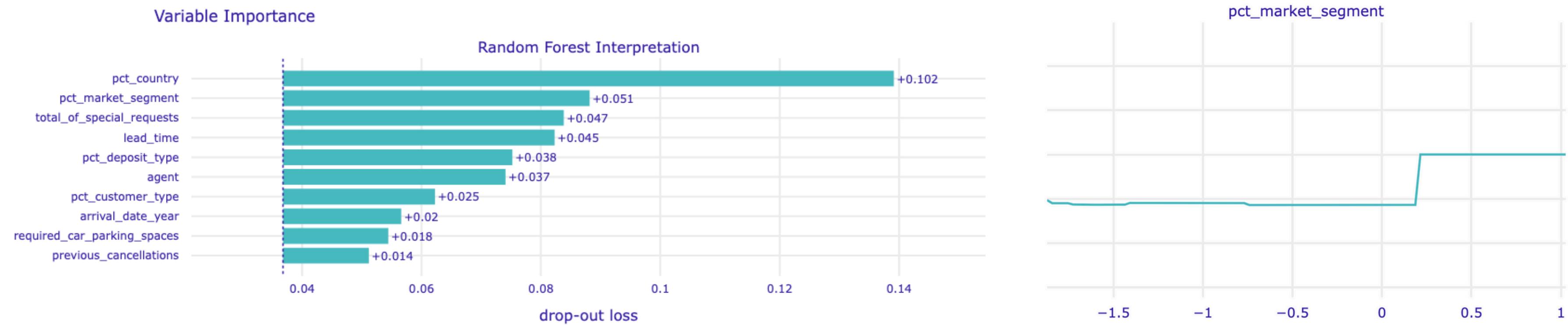
Insight

- Portugal (PRT) memiliki jumlah pembatalan tertinggi dari negara lain.
- Inggris (GBR) dan Jerman (DEU) juga menunjukkan jumlah rasio pembatalan yang sehat (antara sukses vs batalnya banyak yang sukses).

Business Recommendations

- Pertimbangkan untuk meminta deposit atau menerapkan kebijakan pembatalan yang lebih ketat untuk tamu dari negara tingkat pembatalan tinggi (Seperti Portugal)
- Tawarkan fleksibilitas pembatalan khusus untuk jauh hari, dan penalti untuk yang mepet.

Feature Importance (Market Segment)

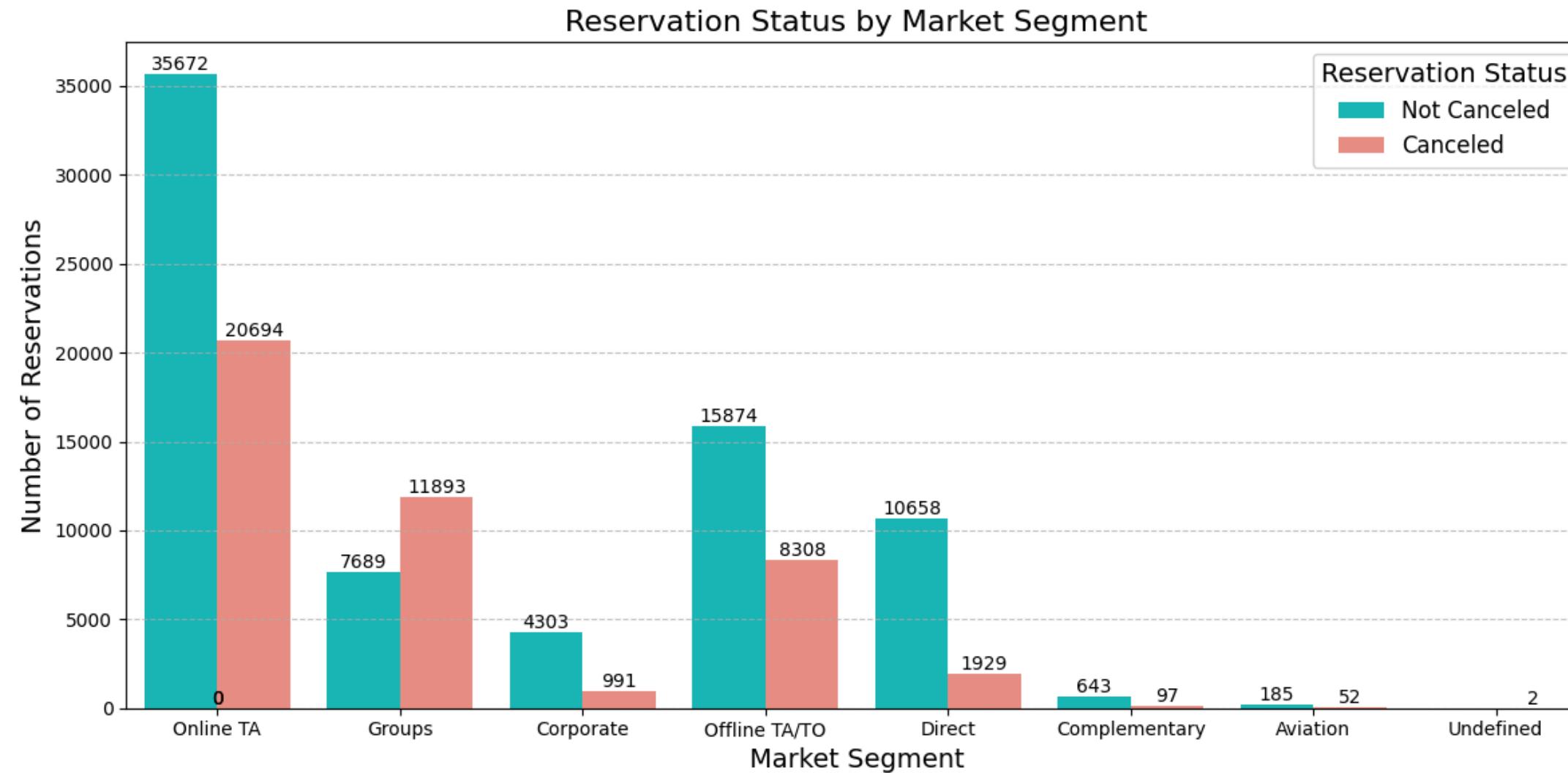


Insight

- Segmentasi pelanggan ternyata berpengaruh terhadap keputusan model. Siapa yang memesan berpengaruh terhadap pembatalan

- Selanjutnya akan dilakukan analisis lebih lanjut terhadap hubungan masing-masing market segment dengan kesuksesan pemesanan (dibatalkan atau tidaknya)

Feature Importance (Cont. Market Segment)



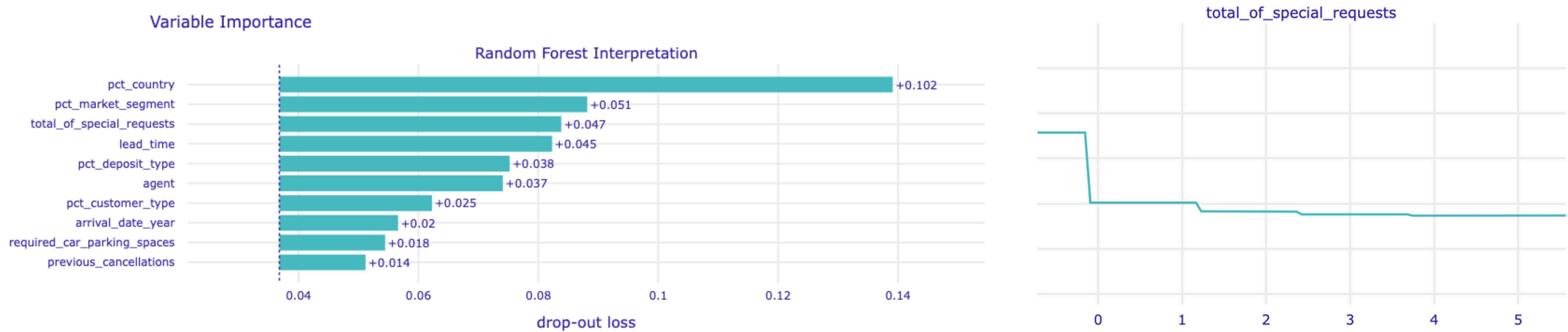
Insight

- Groups memiliki rasio tingkat pembatalan yang paling signifikan
- Online TA (Travel Agents) memiliki volume pemesanan tertinggi, namun tingkat pembatalannya juga masih cukup tinggi
- Direct: Reservasi yang dilakukan secara langsung memiliki tingkat pembatalan yang paling sehat.

Business Recommendations

- Menerapkan kebijakan deposit atau menerapkan kebijakan pembatalan yang lebih ketat, terutama bagi market segment yang rasio pembatalannya tinggi
- Mendorong pelanggan untuk memesan secara langsung (Direct). Bisa dengan menawarkan promosi eksklusif atau loyalty program.

Feature Importance (Total of Special Request)



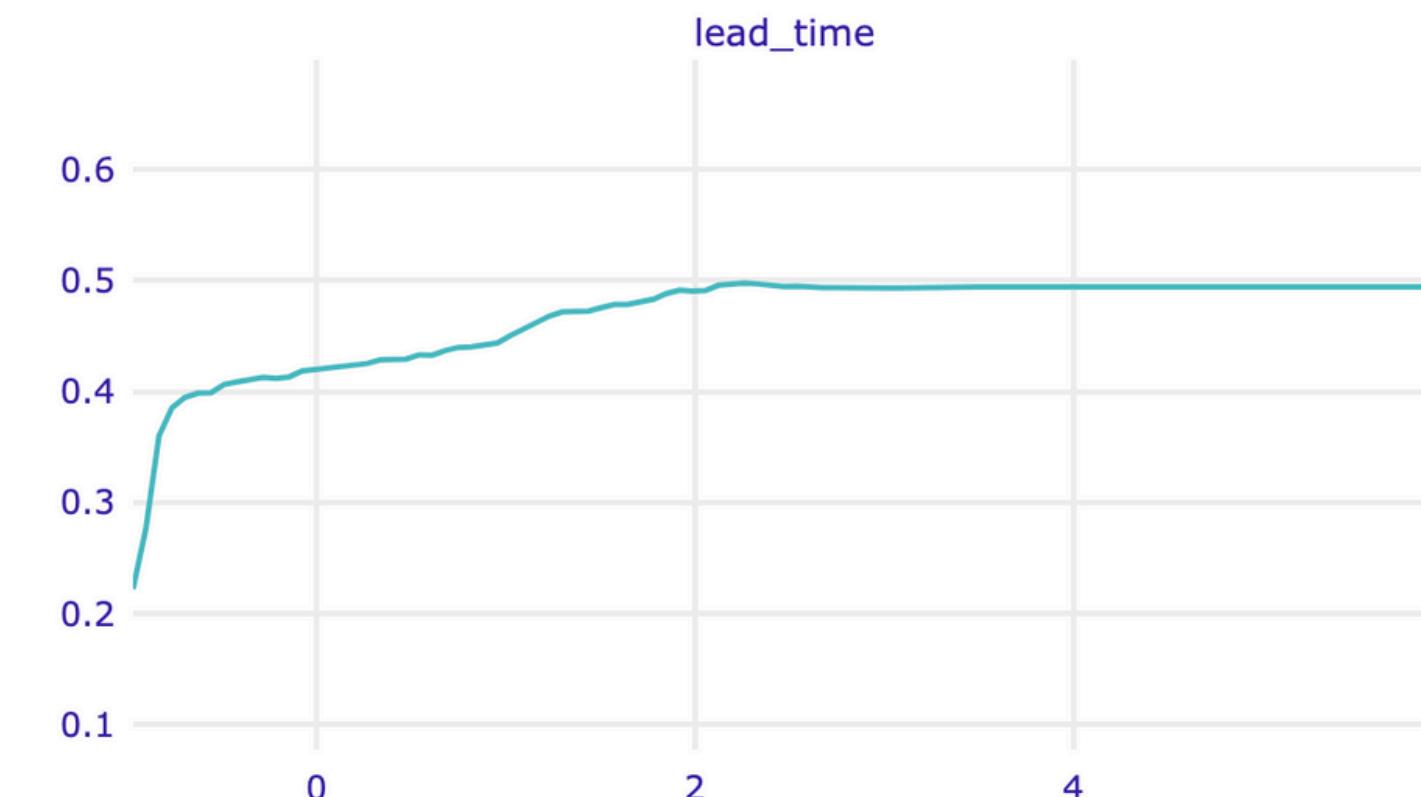
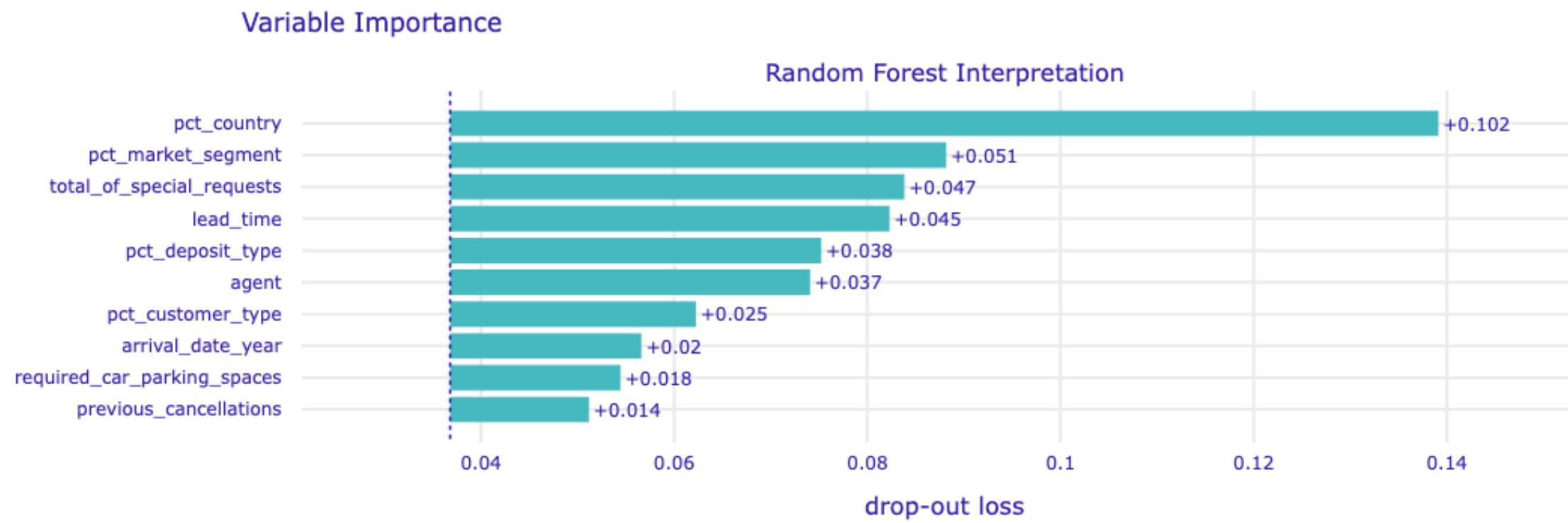
Insight

- Menurut insight saya, mungkin tamu dengan banyak special request cenderung lebih berkomitmen terhadap pesanannya.
- Plot juga membuktikan bahwa semakin tinggi special request, maka ada kecenderungan penurunan kemungkinan pembatalan

Business Recommendations

- Akomodasi special request tamu dengan lebih baik untuk mengurangi kemungkinan pembatalan, dan diinfokan di advertisement
- Mengiklankan bahwa perusahaan menerima opsi special request banyak dan memastikan bahwa permintaan tersebut dapat dipenuhi, sehingga pemesan yang memesan lebih mempertimbangkan special request dan lebih commit terhadap pesanannya.

Feature Importance (Lead Time)



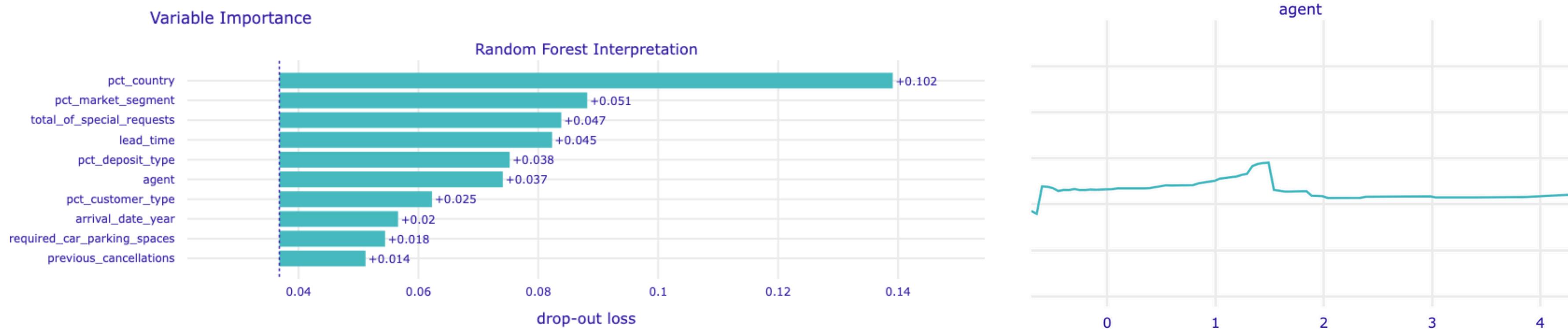
Insight

Jika dilihat dari plot, pemesanan dengan lead time yang makin panjang memiliki kemungkinan untuk melakukan pembatalan yang meningkat.

Business Recommendations

- Kebijakan pembayaran di muka atau deposit untuk pemesanan yang dilakukan jauh hari sebelumnya
- Menawarkan diskon atau insentif untuk pemesanan yang dilakukan dalam waktu dekat dengan tanggal kedatangan

Feature Importance (Agent)

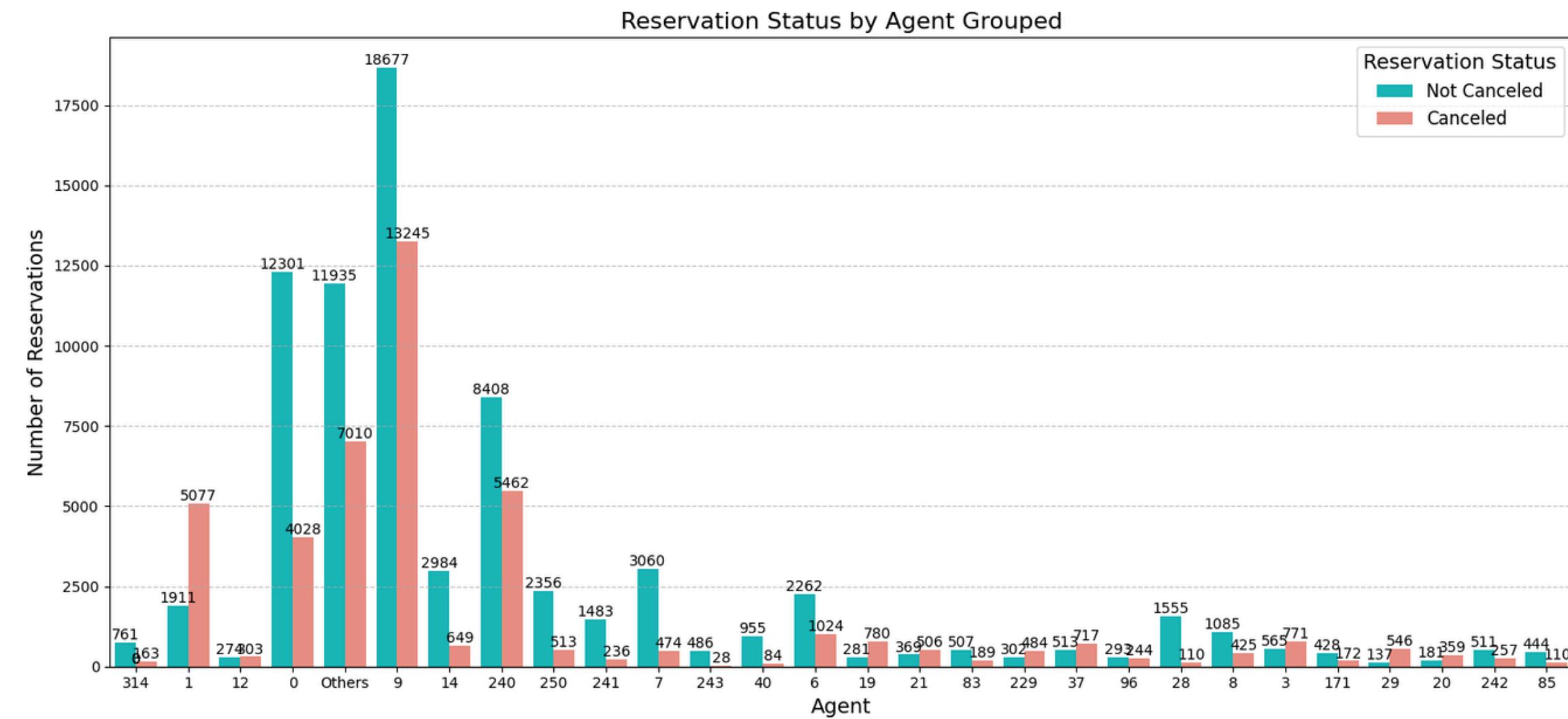


Insight

- Beberapa Agent memiliki tingkat pembatalan yang lebih tinggi dibandingkan yang lain.

- Karena di luar ranah perusahaan, identifikasi Agent yang memiliki tingkat pembatalan tinggi & berkolaborasi dengan Agent tersebut untuk memahami penyebab pembatalan dan mencari solusi untuk menguranginya
- Kita akan identifikasi hubungan Agent terhadap kesuksesan pesanan di slide selanjutnya.

Feature Importance (Cont. Agent)



Insight

- Agent 9 & 1 memiliki jumlah reservasi tertinggi namun tingkat pembatalannya juga tinggi
- Agent seperti 7, 250, dll dengan reservasi lebih rendah memiliki tingkat pembatalan yang relatif lebih rendah

Business Recommendations

- Berkolaborasi dengan Agent 9 & 1 (mengingat tingginya pembatalan), untuk mencari tahu penyebab mayoritas pembatalan.
- Menawarkan insentif kepada Agent seperti 9 & 1 untuk pemesanan yang sukses.

Rangkuman Business Recommendation (I)



Buat Promosi & Penawaran

- Buat paket bundling 2 malam akhir pekan dengan diskon
- Buat diskon progresif untuk menginap >3 malam
- Lakukan promosi pada bulan-bulan dengan ADR menurun
- Pada bulan dengan tingkat pembatalan tinggi (Januari, Februari, Agustus), buat insentif untuk tamu yang tidak membatalkan.
- Tingkatkan pendapatan pada bulan Mei, Juli, dan Agustus

Rangkuman Business Recommendation (2)



Buat Program Loyalitas

- Metode point & rewards untuk meningkatkan tingkat kunjungan tamu berulang
- Beri updates ke customer melalui newsletter terkait penawaran atau promo khusus

Rangkuman Business Recommendation (3)



Evaluasi Kebijakan Pembatalan

- Pertimbangkan biaya pembatalan atau pembayaran DP untuk mengurangi pembatalan dan No-Show.
- Terapkan kebijakan pembatalan yang lebih ketat untuk pemesanan dengan Lead Time yang panjang
- Re-evaluate kebijakan Non Refund. Instead of Non Refund, pertimbangkan insentif khusus bagi tamu yang tidak membatalkan pemesanan.

Rangkuman Business Recommendation (4)



Fasilitas dan Layanan

- Akomodasi special request pelanggan dengan lebih baik dan promosikan penerimaan special request.
- Kumpulkan dan analisa special request pelanggan untuk menyesuaikan penawaran dan trend.
- Manfaatkan ruang parkir (jika terlalu luas) menjadi fasilitas lain seperti taman, area rekreasi, dll.
- Fokus pada kualitas sarapan. Mayoritas customer -> Bed & Breakfast

Rangkuman Business Recommendation (5)



Kolaborasi dengan Agent

- Kolaborasi dengan tingkat pembatalan tinggi untuk memahami penyebab pembatalan dan mencari solusinya.
- Tawarkan insentif kepada agen dengan tingkat pembatalan yang tinggi untuk meningkatkan pemesanan yang sukses.