

**Tugas Besar**  
**IF5100 - Pemrograman untuk Data Analitik**



**Kelompok 10**

13522101	Abdullah Mubarak
23525039	Viody Alfaridzi
23525058	Iskandar Muda Rizky Parlambang

<b>1. Dataset Selection.....</b>	<b>2</b>
<b>2. Data Understanding.....</b>	<b>2</b>
2.1. Data Overview.....	2
2.2. Missing Value.....	2
2.3. Skewness.....	3
2.4. Distribution.....	4
2.4.1. Distribution of Numerical Features.....	4
2.4.2. Distribution of Categorical Features.....	5
2.5. Correlation.....	6
2.5.1. Correlation of Numerical Features.....	6
2.5.1. Correlation Between Features and Target Variable.....	6
2.6. Deeper Visualization.....	7
2.6.1. Cancellation Rate and Count by Hotel Type.....	7
2.6.2. Monthly Arrival by Hotel Type.....	7
2.6.3. Correlation Between is_cancelled with days_before_arrival.....	8
2.6.4. Cancellation Rate by Deposit Type.....	8
2.6.4. Cancellation Rate and Market Segment.....	9
2.6.5. The Effect of Room Mismatch to Cancellation Rate.....	9
<b>3. Data Preparation.....</b>	<b>9</b>
3.1. Data Cleaning and Quality.....	9
3.2. Handling Missing Data.....	10
3.2.1. For Numerical Missing Value.....	10
3.2.2. For Categorical Missing Value.....	11
3.3. Dealing With Outlier.....	11
3.4. Feature Engineering.....	12
3.5. Compile Preprocessing Pipeline.....	12
<b>4. Inference.....</b>	<b>13</b>
4.1. Modeling and Validation.....	13
4.2. Model Evaluation.....	14
<b>5. Links and References.....</b>	<b>14</b>
5.1. Important Links.....	14
5.2. References.....	14
<b>6. Contribution.....</b>	<b>15</b>
6.1. Contribution table.....	15

## 1. Dataset Selection

Dataset yang dipilih adalah Dataset Hotel Booking Demand yang terdapat dari kaggle. Dataset tersebut berisi informasi booking untuk city hotel dan resort hotel yang memiliki 32 kolom. Informasi yang ada diantaranya adalah tanggal booking dilakukan, durasi pemakaian, jumlah orang dewasa, jumlah anak kecil, dan sebagainya. Dataset ini memiliki kolom yang banyak sehingga banyak juga hal yang dapat dianalisis pada tahap Exploratory Data Analysis (EDA). Data-data yang terdapat pada dataset tersebut yang sekilas terlihat bersih juga menjadi alasan pemilihan dataset ini. Rencana selanjutnya adalah data-data ini akan digunakan untuk membuat model untuk memprediksi booking akan dibatalkan atau tidak dengan kolom target adalah `is_canceled`.

## 2. Data Understanding

### 2.1. Data Overview

Dapat terlihat pada gambar dibawah, Dataset ini memiliki 32 kolom dan 119390 baris dengan 4 tipe data `float`, 16 tipe data `int`, dan 12 tipe data `object`. Dapat dilihat dari statistik kolom numerik pada dataset ini, nilai minimal kolom `adults` adalah 0. Hal tersebut seharusnya mustahil terjadi dan kemungkinan pemesanan tersebut merupakan pemesanan palsu. Oleh karena itu, diperlukan filtering untuk memastikan jumlah `adults` lebih dari 0.

Data columns (total 32 columns):		
#	Column	Non-Null Count Dtype
0	hotel	95512 non-null object
1	is_canceled	95512 non-null int64
2	lead_time	95512 non-null int64
3	arrival_date_year	95512 non-null int64
4	arrival_date_month	95512 non-null object
5	arrival_date_week_number	95512 non-null int64
6	arrival_date_day_of_month	95512 non-null int64
7	stays_in_weekend_nights	95512 non-null int64
8	stays_in_week_nights	95512 non-null int64
9	adults	95512 non-null int64
10	children	95508 non-null float64
11	babies	95512 non-null int64
12	meal	95512 non-null object
13	country	95048 non-null object
14	market_segment	95512 non-null object
15	distribution_channel	95512 non-null object
16	is_repeated_guest	95512 non-null int64
17	previous_cancellations	95512 non-null int64
18	previous_bookings_not_canceled	95512 non-null int64
19	reserved_room_type	95512 non-null object
20	assigned_room_type	95512 non-null object
21	booking_changes	95512 non-null int64
22	deposit_type	95512 non-null object
23	agent	81726 non-null float64
24	company	5581 non-null float64
25	days_in_waiting_list	95512 non-null int64
26	customer_type	95512 non-null object
27	adr	95512 non-null float64
28	required_car_parking_spaces	95512 non-null int64
29	total_of_special_requests	95512 non-null int64
30	reservation_status	95512 non-null object
31	reservation_status_date	95512 non-null object
dtypes: float64(4), int64(16), object(12)		

	count	mean	std	min	25%	50%	75%	max
lead_time	95512.0	102.219732	106.932140	0.00	17.0	66.0	158.0000	737.0
arrival_date_year	95512.0	2015.945693	0.635090	2015.00	2016.0	2016.0	2016.0000	2017.0
arrival_date_week_number	95512.0	28.099694	14.667629	1.00	15.0	30.0	41.0000	53.0
arrival_date_day_of_month	95512.0	15.652180	8.762123	1.00	8.0	16.0	23.0000	31.0
stays_in_weekend_nights	95512.0	0.906399	0.997957	0.00	0.0	1.0	2.0000	19.0
stays_in_week_nights	95512.0	2.464874	1.899039	0.00	1.0	2.0	3.0000	50.0
adults	95512.0	1.842952	0.599374	0.00	2.0	2.0	2.0000	55.0
children	95508.0	0.094390	0.380700	0.00	0.0	0.0	0.0000	10.0
babies	95512.0	0.007989	0.099359	0.00	0.0	0.0	0.0000	10.0
is_repeated_guest	95512.0	0.031253	0.174001	0.00	0.0	0.0	0.0000	1.0
previous_cancellations	95512.0	0.105526	0.937579	0.00	0.0	0.0	0.0000	26.0
previous_bookings_not_canceled	95512.0	0.124477	1.354917	0.00	0.0	0.0	0.0000	61.0
booking_changes	95512.0	0.204613	0.622661	0.00	0.0	0.0	0.0000	21.0
agent	81726.0	83.691616	107.176901	1.00	9.0	14.0	195.0000	510.0
company	5587.0	180.781457	126.356734	6.00	51.0	174.0	242.0000	516.0
days_in_waiting_list	95512.0	2.866949	19.552508	0.00	0.0	0.0	0.0000	391.0
adr	95512.0	94.949131	47.082480	-6.38	65.0	89.0	116.8275	5400.0
required_car_parking_spaces	95512.0	0.061333	0.241029	0.00	0.0	0.0	0.0000	3.0
total_of_special_requests	95512.0	0.512585	0.759502	0.00	0.0	0.0	1.0000	5.0

### 2.2. Missing Value

Dapat terlihat pada gambar dibawah, terdapat 4 kolom dengan missing value. Kolom `company` dengan missing value sangat besar (94.15%) sehingga informasi perusahaan pemesan hampir tidak tersedia dan kemungkinan tidak berguna untuk pemodelan. Kolom `agent` memiliki missing 14.43% cukup signifikan sehingga perlu dipertimbangkan strategi imputasi atau penghapusannya. Kolom `country` hanya memiliki 0.49% missing sehingga dapat diimputasi

tanpa mengganggu kualitas data. Kolom lainnya tidak memiliki missing value sehingga aman digunakan langsung.

	missing_count	missing_percent
company	112221	94.31
agent	16263	13.67
country	478	0.40
children	4	0.00

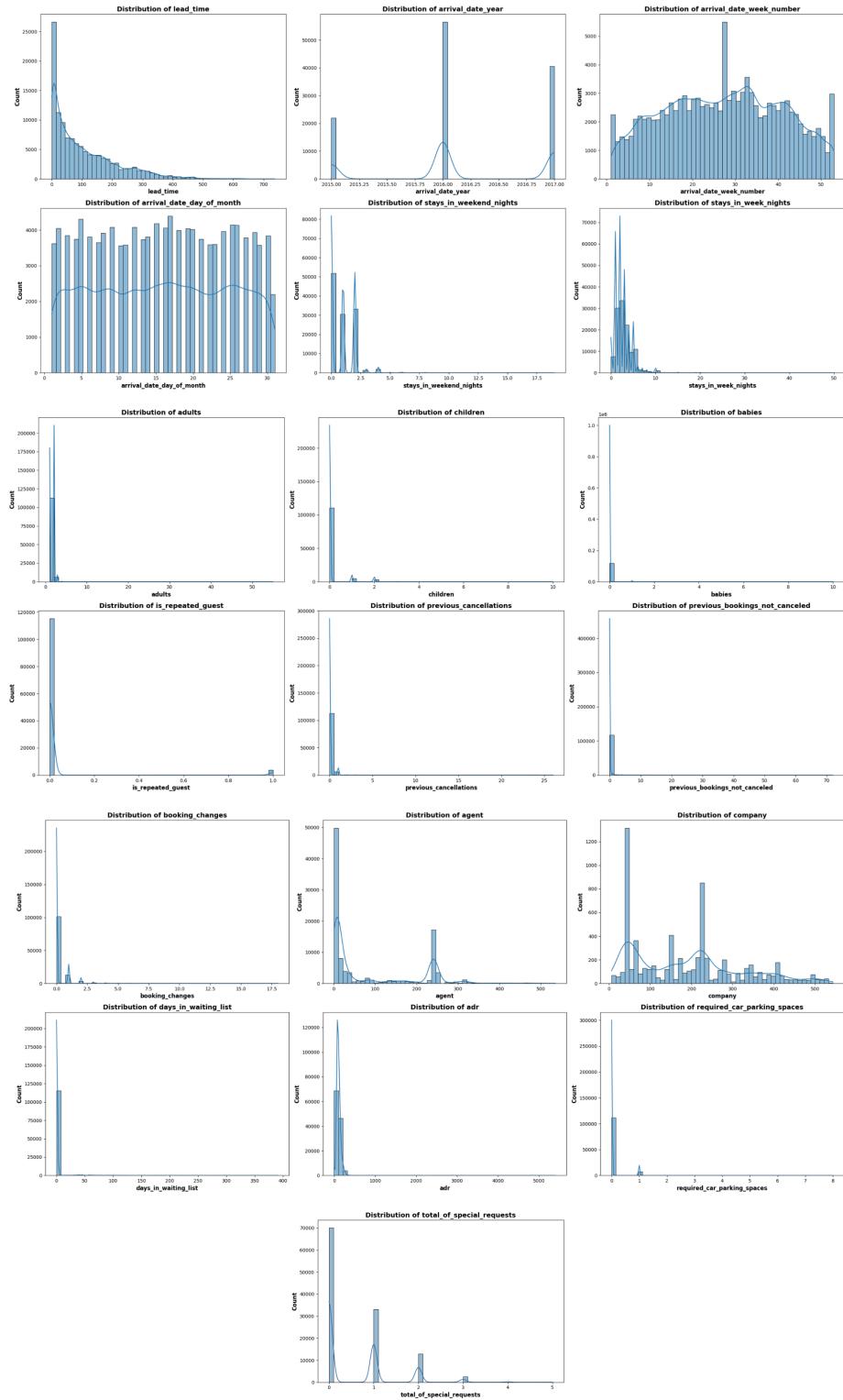
### 2.3. Skewness

Dapat terlihat pada gambar dibawah, mayoritas fitur numerik memiliki nilai skewness yang tinggi. Hal tersebut mengindikasikan distribusi data yang sangat miring dengan keberadaan outlier yang signifikan. Oleh karena itu, penanganan missing values lebih baik dilakukan menggunakan metode imputasi median. Pendekatan ini dipilih karena median lebih *robust* sehingga mampu memberikan estimasi yang lebih representatif dibandingkan dengan mean.

lead_time	1.359995
arrival_date_year	0.045149
arrival_date_week_number	0.153469
arrival_date_day_of_month	0.015691
stays_in_weekend_nights	1.474751
stays_in_week_nights	3.046951
adults	20.734764
children	4.373227
babies	26.924032
is_repeated_guest	5.387995
previous_cancellations	22.217416
previous_bookings_not_canceled	22.585756
booking_changes	6.258030
agent	1.043635
company	0.574965
days_in_waiting_list	10.747626
adr	15.770103
required_car_parking_spaces	3.716997
total_of_special_requests	1.439837

## 2.4. Distribution

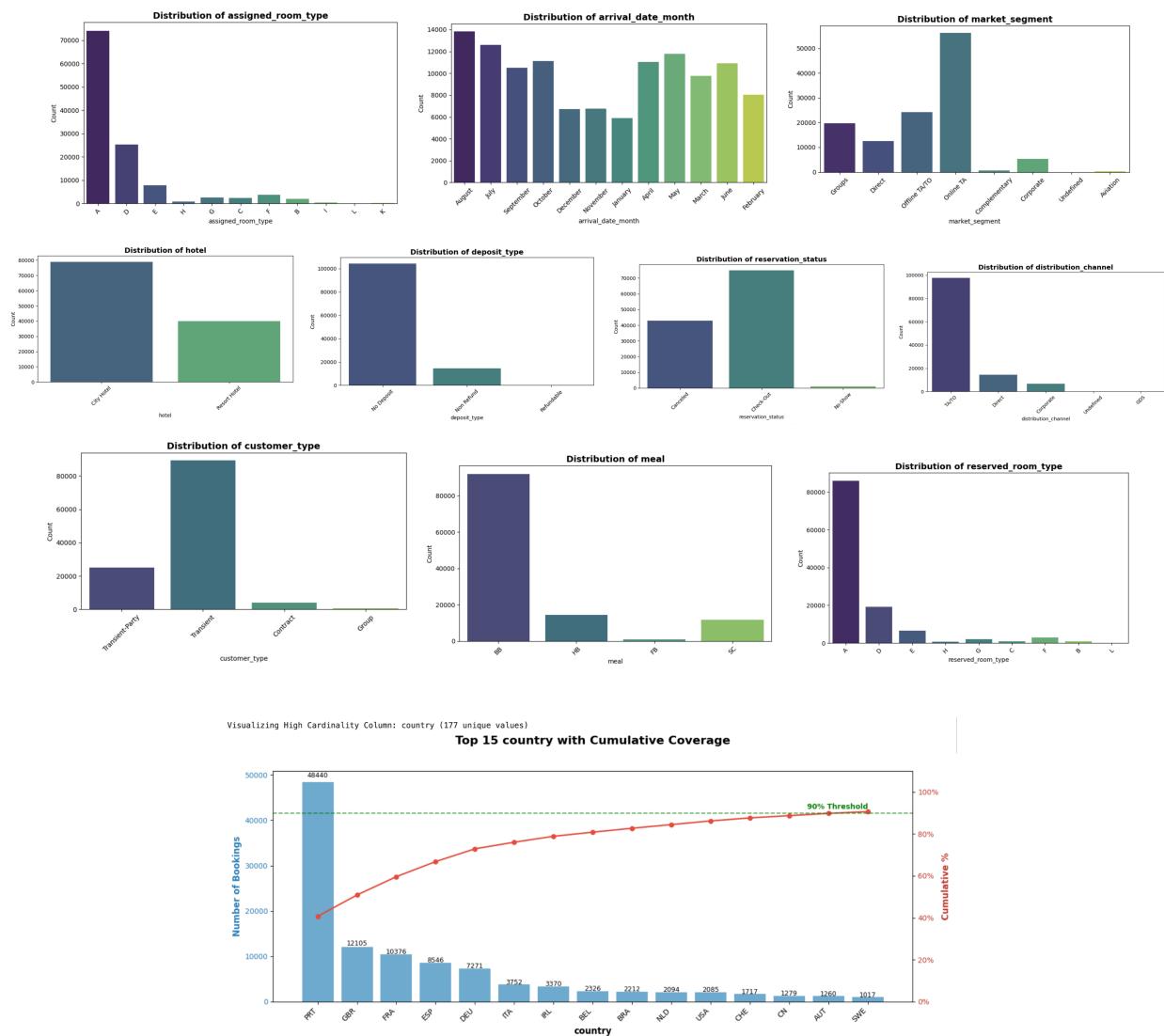
### 2.4.1. Distribution of Numerical Features

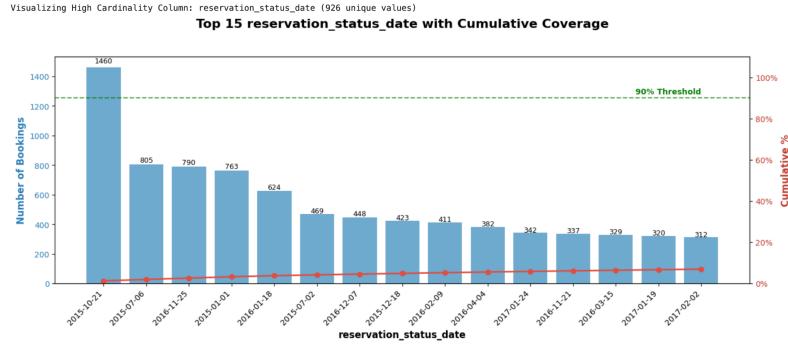


## 2.4.2. Distribution of Categorical Features

Pada distribusi kolom Country didapatkan bahwa PRT (Portugal) mendominasi grafik secara ekstrem (sekitar 40-50% data). Garis merah (Cumulative %) menunjukkan bahwa Top 10-15 negara saja sudah mencakup >90% dari total data. Dataset ini memiliki total sekitar 170+ negara. Grafik menunjukkan bahwa negara ke-16 sampai ke-177 memiliki frekuensi yang sangat kecil (bar-nya hampir tidak terlihat).

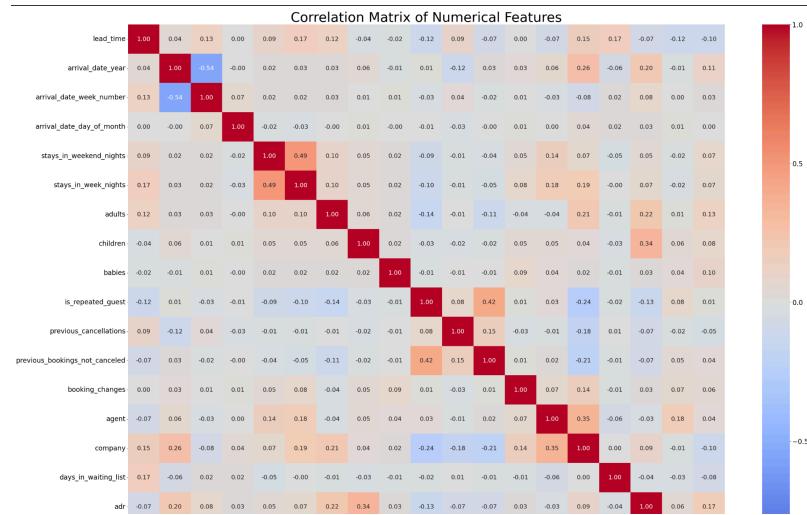
Membeliarkan ratusan kolom negara kecil ini (jika di One-Hot Encode) hanya akan menambah noise dan beban komputasi (*Curse of Dimensionality*) tanpa memberikan pola yang signifikan. Oleh karena itu, akan dilakukan penggabungan semua negara di luar Top 10/15 menjadi satu kategori "Other".





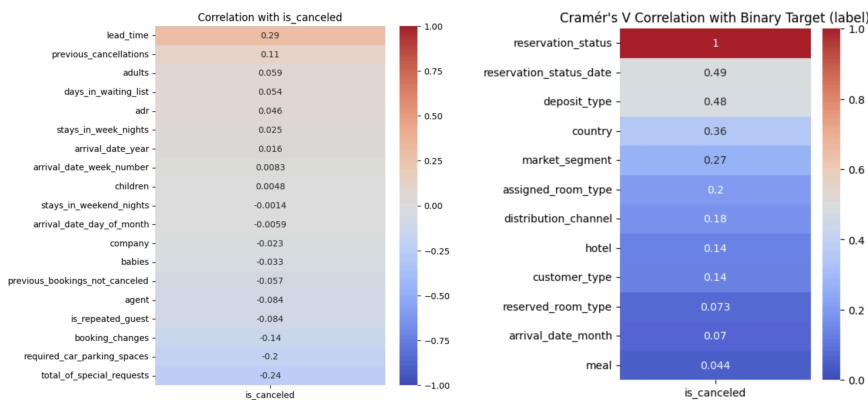
## 2.5. Correlation

### 2.5.1. Correlation of Numerical Features



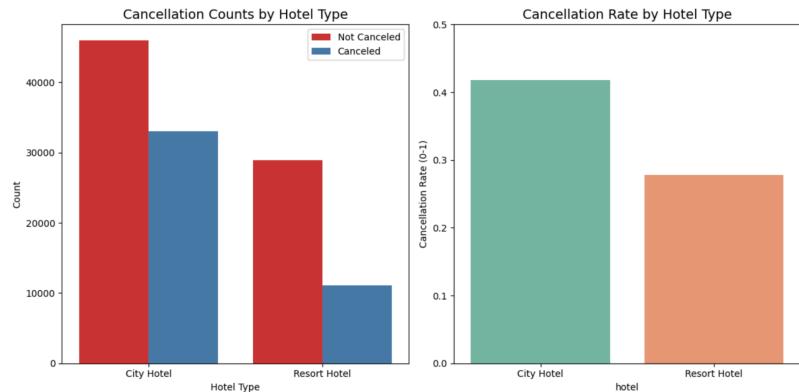
### 2.5.1. Correlation Between Features and Target Variable

Korelasi numerical features dengan target dapat dilihat pada gambar kiri dan korelasi categorical features dengan target dapat dilihat pada gambar kanan.



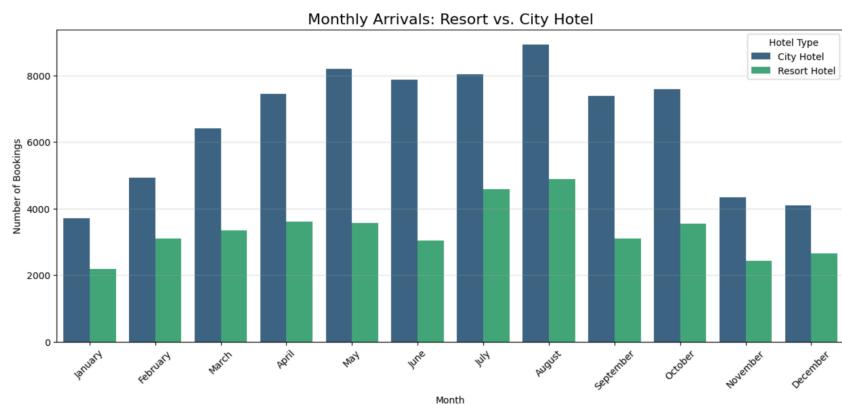
## 2.6. Deeper Visualization

### 2.6.1. Cancellation Rate and Count by Hotel Type



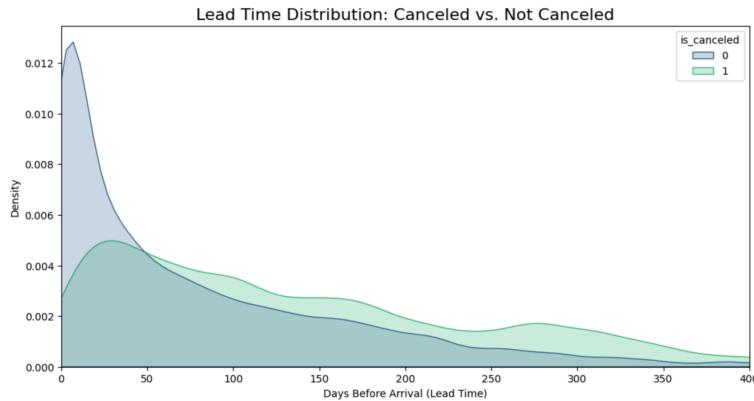
Terdapat perbedaan perilaku yang signifikan antara kedua jenis hotel. City Hotel memiliki volume pemesanan yang jauh lebih tinggi namun juga memiliki tingkat pembatalan (cancellation rate) yang lebih besar (sekitar 42%). Sebaliknya, Resort Hotel cenderung lebih stabil dengan tingkat pembatalan yang lebih rendah (sekitar 28%).

### 2.6.2. Monthly Arrival by Hotel Type



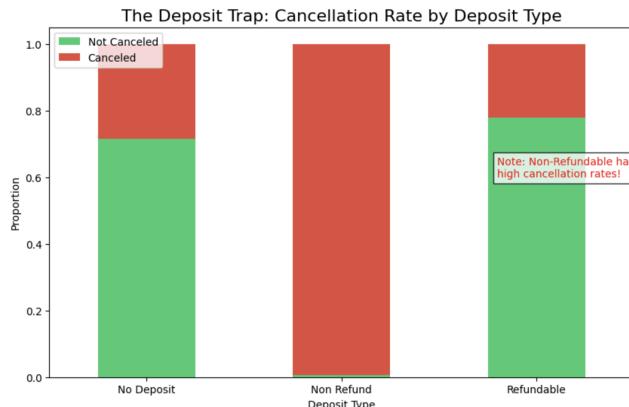
Pola musiman (seasonality) terlihat jelas pada Resort Hotel, di mana kunjungan melonjak tajam pada bulan Juli dan Agustus (musim panas). Sementara itu, City Hotel memiliki permintaan yang relatif tinggi dan konsisten sepanjang tahun, dengan sedikit penurunan di musim dingin, menunjukkan bahwa hotel kota tidak hanya bergantung pada turis liburan tetapi juga pebisnis.

### 2.6.3. Correlation Between is\_cancelled with days\_before\_arrival



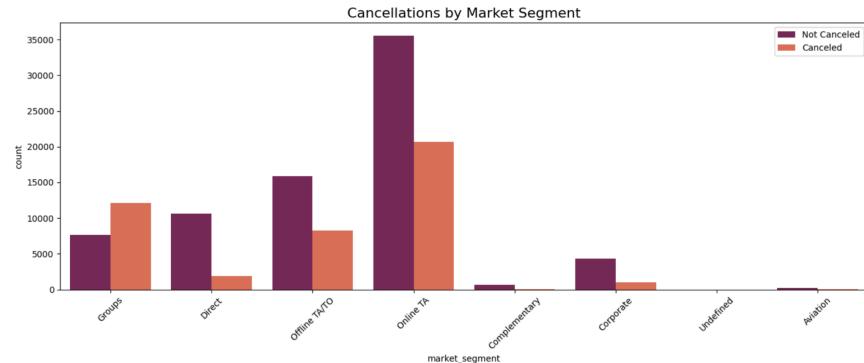
Semakin lama jarak waktu pemesanan (lead time), semakin tinggi risiko pembatalan. Grafik menunjukkan bahwa tamu yang memesan dalam waktu singkat (kurva hijau tinggi di kiri) cenderung tidak membatalkan pesanan. Sebaliknya, pemesanan yang dilakukan jauh-jauh hari (kurva biru melebar ke kanan) memiliki probabilitas pembatalan yang jauh lebih tinggi karena rencana tamu lebih mungkin berubah.

### 2.6.4. Cancellation Rate by Deposit Type



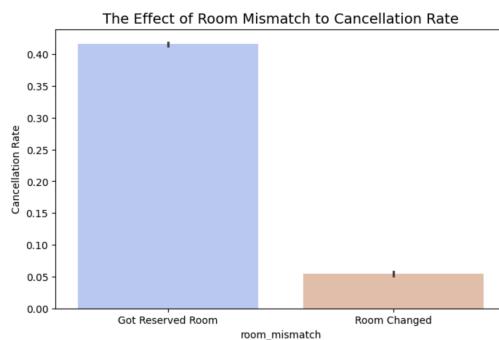
Hal ini merupakan anomali terkenal dalam dataset ini. Tipe deposit "Non Refund" justru memiliki tingkat pembatalan yang hampir 100%. Ini berlawanan dengan intuisi, namun biasanya terjadi karena jenis deposit ini sering diterapkan pada pemesanan grup yang berisiko tinggi atau pemesanan fiktif yang akhirnya dibatalkan oleh sistem atau agen.

## 2.6.5. Cancellation Rate and Market Segment



Segmen "Online TA" (Travel Agent) adalah penyumbang volume terbesar namun memiliki jumlah pembatalan yang juga masif. Segmen "Groups" sangat berisiko dengan rasio pembatalan yang tinggi dibandingkan pemesanan suksesnya. Di sisi lain, segmen "Direct" (tamu yang memesan langsung ke hotel) dan "Corporate" adalah segmen paling setia dengan tingkat pembatalan yang sangat minim.

## 2.6.6. The Effect of Room Mismatch to Cancellation Rate



Grafik ini menunjukkan fenomena yang kontraintuitif namun menarik. Tamu yang mendapatkan kamar berbeda dari yang mereka pesan ("Room Changed") justru memiliki tingkat pembatalan yang sangat rendah (hanya sekitar 5%). Hal ini mungkin terjadi karena dalam operasional hotel, perubahan tipe kamar sering kali berarti upgrade gratis ke kamar yang lebih mahal/luas. Tamu yang menerima upgrade tentu merasa diuntungkan dan hampir tidak mungkin membatalkan pesanan, berbeda dengan tamu yang mendapatkan kamar sesuai pesanan awal ("Got Reserved Room") yang memiliki tingkat risiko pembatalan yang relatif normal.

## 3. Data Preparation

### 3.1. Data Cleaning and Quality

Pada tahap ini, dilakukan penghapusan row data yang tidak valid di mana jumlah orang dewasa (adults) bernilai 0. Selain itu, kualitas data juga diperbaiki menggunakan InitialCleaner. Proses ini meliputi mengganti kategori "Undefined" menjadi "SC" pada kolom Meal dan kolom negara disederhanakan dengan hanya mempertahankan 15 negara teratas berdasarkan frekuensi, sementara sisanya dikelompokkan ke dalam kategori "Other".

### **3.2. Handling Missing Data**

Penanganan missing value numerik dilakukan dengan dua strategi imputasi berbeda sesuai dengan karakteristik kolomnya menggunakan NumMissingValueHandler. Kolom-kolom yang secara logika harus diisi 0 jika kosong, yaitu agent, company, children, babies, required\_car\_parking\_spaces, dan total\_of\_special\_requests. Dari penjelasan dataset, diketahui bahwa missing value pada agent berarti booking dilakukan langsung. Sementara itu, kolom numerik sisanya yang cenderung memiliki distribusi skewed diisi menggunakan nilai median agar lebih tahan terhadap outlier. Penanganan missing value kategorikal disesuaikan dengan jenis atributnya menggunakan CatMissingValueHandler. Khusus untuk kolom country, nilai yang kosong dikategorikan sebagai kelas baru, yaitu “Other”. Sedangkan untuk kolom kategori lainnya, missing value akan diimputasi menggunakan nilai modus dari masing-masing kolom.

### **3.3. Dealing With Outlier**

Penanganan outlier dilakukan menggunakan metode heuristik melalui OutlierHandler dimana batas nilai maksimum ditentukan berdasarkan logika untuk membatasi nilai ekstrim yang tidak wajar. Nilai yang melebihi ambang batas tertentu, seperti jumlah tamu dewasa di atas 10 orang atau jumlah bayi di atas 5 akan langsung dibatasi menjadi nilai batas maksimum yang telah ditetapkan.

### **3.4. Feature Engineering**

Pada tahap ini, langkah pertama yang dilakukan ada remove useless attribute dengan tujuan membersihkan dataset dari atribut yang tidak memberikan nilai tambah atau berisiko merusak validitas model. Kolom company dihapus karena memiliki terlalu banyak missing values, sedangkan kolom reservation\_status dan reservation\_status\_date juga dihapus untuk mencegah terjadinya data leakage terhadap target prediksi. Langkah kedua yang dilakukan pada tahap ini adalah adding new feature. Fitur-fitur baru diekstraksi untuk memperkaya informasi yang dapat dipelajari oleh model. Penambahan fitur ini mencakup total\_guests, total\_nights, room\_mismatch (apakah ruangan yang didapat sesuai booking), dan is\_family (apakah membawa keluarga).

### **3.5. Compile Preprocessing Pipeline**

Seluruh tahapan preprocessing akan disatukan ke dalam satu alur kerja otomatis (final\_pipeline). Pipeline ini mengintegrasikan pembersihan awal, seleksi fitur, serta transformasi spesifik di mana data numerik dinormalisasi distribusinya menggunakan PowerTransformer dan disetarakan skalanya dengan RobustScaler, sementara data kategorikal diubah menjadi format numerik melalui OneHotEncoder.

## **4. Inference**

### **4.1. Modeling and Validation**

Pada tahap ini, kami mengembangkan tiga jenis model machine learning dengan pendekatan algoritma yang berbeda, yaitu Logistic Regression, Random Forest, dan XGBoost. Seluruh atribut yang telah diproses digunakan untuk memprediksi target, yaitu kemungkinan pembatalan (is\_canceled).

Prediksi ini dalam skenario dunia nyata sangat berguna bagi manajemen hotel untuk mengoptimalkan strategi overbooking dan meminimalisir pendapatan yang hilang akibat kamar kosong yang tidak terduga.

Untuk memastikan performa yang optimal, setiap model melalui proses Hyperparameter Tuning otomatis menggunakan Optuna dengan skema 100 trials. Proses ini bertujuan mencari kombinasi parameter terbaik guna memaksimalkan akurasi pada data validasi. Hasil tuning menunjukkan bahwa model ensemble (Random Forest dan XGBoost) membutuhkan kompleksitas yang lebih tinggi untuk mempelajari karakteristik data secara efektif dibandingkan model linear sederhana. Berikut adalah hasil parameter terbaik untuk setiap model dari Hyperparameter Tuning:

1. Logistic Regression:
  - max\_iter=1000
  - random\_state=42
  - C=6.46813777925143
  - solver='lbfgs'
2. Random Forest:
  - n\_estimators=291
  - max\_depth=35
  - min\_samples\_split=2
  - min\_samples\_leaf=1
  - random\_state=42
  - n\_jobs=-1
3. XGBoost:
  - use\_label\_encoder=False
  - eval\_metric='logloss'
  - random\_state=42
  - n\_jobs=-1
  - early\_stopping\_rounds=30
  - n\_estimators=219
  - max\_depth=9
  - learning\_rate= 0.16342802729741931
  - subsample=0.9440447396399663
  - colsample\_bytree=0.7422161132970942
  - gamma=0.13247505415284713
  - reg\_alpha=0.380620896437301
  - reg\_lambda=0.8982764996493284

## 4.2. Model Evaluation

Hasil evaluasi menunjukkan bahwa Random Forest menjadi model terbaik dengan akurasi 91.47% dengan ROC-AUC Score 0.906. Random Forest terlihat lebih unggul dibandingkan XGBoost (91.37%) dan Logistic Regression (84.66%). Selain itu, Random Forest menunjukkan performa yang sangat seimbang dengan F1-Score rata-rata sebesar 0.91. Model ini juga mampu memprediksi dengan baik tamu yang tidak membatalkan (Precision: 91% dan Recall 95%) dan tamu yang membatalkan (Precision: 93% dan Recall 86%). Keunggulan ini sangat krusial mengingat distribusi kelas pada data uji

sebenarnya tidak seimbang, dimana kelas 0 mendominasi (59%) dibandingkan dengan kelas 1 (41%). Meskipun menghadapi ketimpangan data, Random Forest terbukti tidak bias ke kelas mayoritas. Insight ini memberikan kepercayaan tinggi bagi bisnis perhotelan untuk menggunakan model ini dalam pengambilan keputusan karena risiko kesalahan prediksi relatif rendah.

```

Model Evaluation:
Classification Report Logistic Regression:
      precision    recall  f1-score   support
0           0.84     0.91    0.88   11299
1           0.85     0.75    0.80    7736

   accuracy          0.85      --   19035
  macro avg       0.85     0.83    0.84   19035
weighted avg    0.85     0.85    0.84   19035

Classification Report Random Forest:
      precision    recall  f1-score   support
0           0.91     0.95    0.93   11299
1           0.93     0.86    0.89    7736

   accuracy          0.91      --   19035
  macro avg       0.92     0.91    0.91   19035
weighted avg    0.92     0.91    0.91   19035

Classification Report XGBoost:
      precision    recall  f1-score   support
0           0.91     0.94    0.93   11299
1           0.91     0.87    0.89    7736

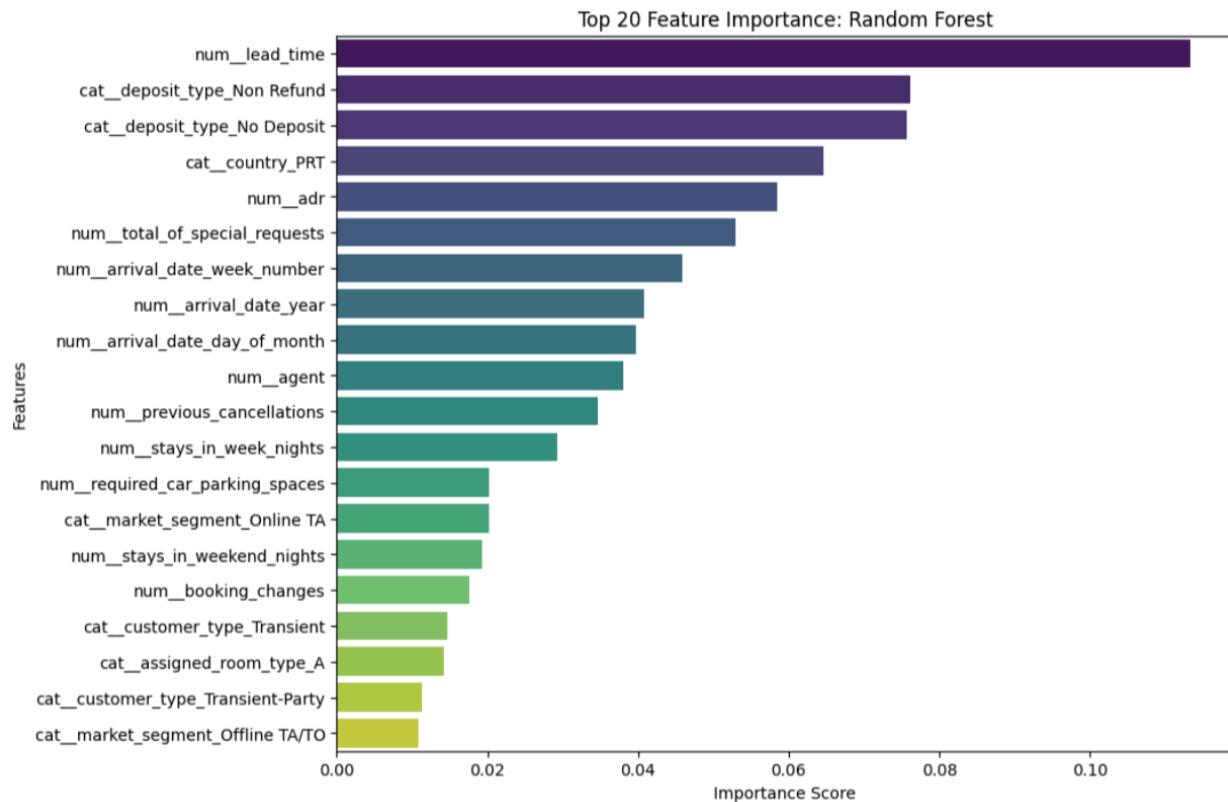
   accuracy          0.91      --   19035
  macro avg       0.91     0.91    0.91   19035
weighted avg    0.91     0.91    0.91   19035

Evaluation Results:
      Model  Accuracy   ROC-AUC
1  Random Forest  0.914683  0.905655
2    XGBoost      0.913738  0.906815
0  Logistic Regression  0.846598  0.831775

```

## 4.2. Feature Importance

Berdasarkan grafik Feature Importance dari model terbaik (Random Forest), terlihat bahwa num\_lead\_time merupakan prediktor paling dominan. Hal ini mengindikasikan bahwa perilaku perencanaan waktu tamu adalah faktor utama penentu pembatalan. Faktor krusial berikutnya berkaitan dengan komitmen finansial yang ditunjukkan oleh tingginya pengaruh deposit\_type. Selain itu, fitur country\_PRT dan adr juga lebih signifikan pengaruhnya dibandingkan fitur operasional lain seperti tipe kamar dan saluran pemesanan.



## 5. Links and References

### 5.1. Important Links

Dataset: <https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand/data>

Github:

### 5.2. References

## 6. Contribution

### 6.1. Contribution table

<b>Nama</b>	Abdullah Mubarak	
<b>NIM</b>	13522101	
<b>No</b>	<b>Kegiatan</b>	<b>Hasil</b>
1	Mencari Dataset	Dataset Fix

2	Melakukan EDA	Hasil EDA
3	Membuat Deeper Visualization	Deeper visualization
4	Menyusun Dokumen Laporan	Dokumen Laporan
5	Membuat Data Cleaning dan Preprocessing	Keseluruhan Pipeline Data

<b>Nama</b>	Viody Alfaridzi	
<b>NIM</b>	23525039	
<b>No</b>	<b>Kegiatan</b>	<b>Hasil</b>
1	Mencari Dataset	Dataset Fix
2	Melakukan Adjustment pada EDA	Adjustment for better visualization and wording
3	Menyusun Dokumen Laporan	Dokumen Laporan

<b>Nama</b>	Iskandar Muda Rizky Parlambang	
<b>NIM</b>	23525058	
<b>No</b>	<b>Kegiatan</b>	<b>Hasil</b>
1	Mencari Dataset	Dataset Fix
2	Melakukan Modeling dengan Hyperparameter Tuning, Evaluation, dan Feature Importance	Model, hasil evaluasi model, dan analisis feature importance
3	Menyusun Dokumen Laporan	Dokumen Laporan