# LAPORAN HOMEWORK EDA

# ANGGOTA KELOMPOK

Celestial Randy

Sonia Epifany Sandah

Oky Hariawan

Risca Naquitasia

Mochamad Choiril Iman

Ahmad Reza

Yehezkiel Novianto A.

# EDA

Exploratory Data Analysis

# STATISTIC DESCRIPTIVE (1)

## Insight:

- Ada beberapa kolom yang kosong atau memiliki null value seperti children, country, agent, company.
- Ada tipe data yang tidak sesuai yaitu children, agent, dan company yaitu float64 harusnya int64 (reservation_status_date yang dapat dijadikan datetime)
- Tidak ditemukan data yang duplikat

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 36 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         118902 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  object
 32  name                            119390 non-null  object
 33  email                           119390 non-null  object
 34  phone-number                    119390 non-null  object
 35  credit_card                     119390 non-null  object
dtypes: float64(4), int64(16), object(16)
memory usage: 32.8+ MB
```

# STATISTIC DESCRIPTIVE (2)

## Insight:

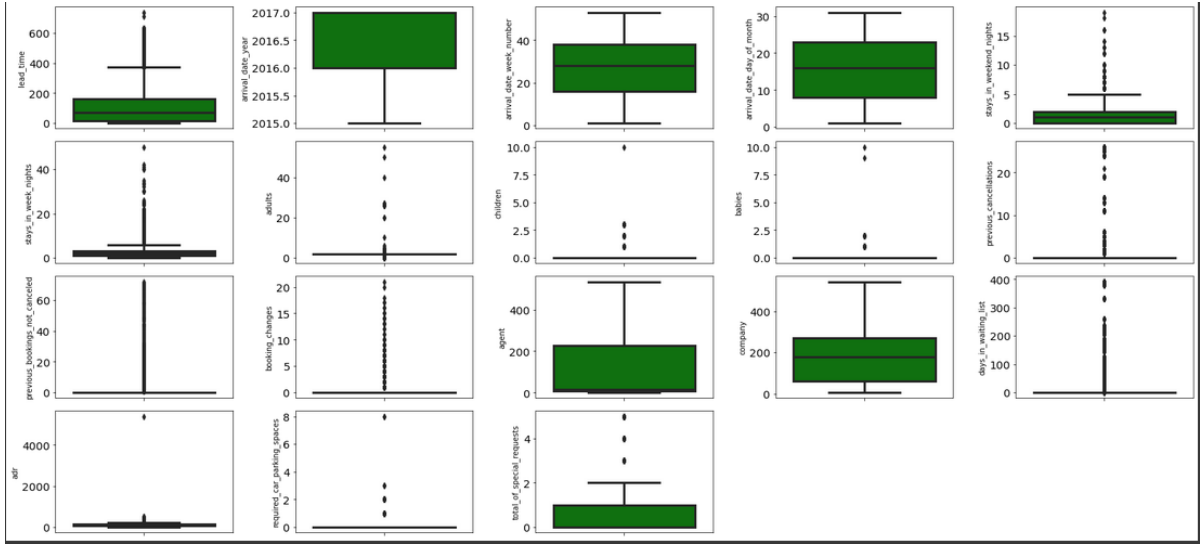
Beberapa kolom previous_cancellations, previous_bookings_not_canceled, booking_changes days_in_waiting_list, required_car_parking_spaces, total_of_special_requests hanya memiliki maximum value (Q4). (IQR) interquartile nya tidak menyebar, hanya di Q4 saja, Hal ini menunjukan bahwa kolom ini hanya berkontribusi pembatalan booking hanya pada beberapa kasus saja. data bagus jika menyebar di Q2-Q3 (besar box ditengah))

| | 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 |
|---|---|---|---|---|---|---|---|---|
| count | 119390.000000 | 119390.000000 | 119390.000000 | 119390.000000 | 119390.000000 | 119390.000000 | 119390.000000 | 119390.000000 |
| mean | 0.370416 | 104.011416 | 2016.156554 | 27.165173 | 15.798241 | 0.927599 | 2.500302 | 1.856403 |
| std | 0.482918 | 106.863097 | 0.707476 | 13.605138 | 8.780829 | 0.998613 | 1.908286 | 0.579261 |
| min | 0.000000 | 0.000000 | 2015.000000 | 1.000000 | 1.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 0.000000 | 18.000000 | 2016.000000 | 16.000000 | 8.000000 | 0.000000 | 1.000000 | 2.000000 |
| 50% | 0.000000 | 69.000000 | 2016.000000 | 28.000000 | 16.000000 | 1.000000 | 2.000000 | 2.000000 |
| 75% | 1.000000 | 160.000000 | 2017.000000 | 38.000000 | 23.000000 | 2.000000 | 3.000000 | 2.000000 |
| max | 1.000000 | 737.000000 | 2017.000000 | 53.000000 | 31.000000 | 19.000000 | 50.000000 | 55.000000 |

# UNIVARIATE ANALYSIS NUMERICAL (BOXPLOT)

## Insight:

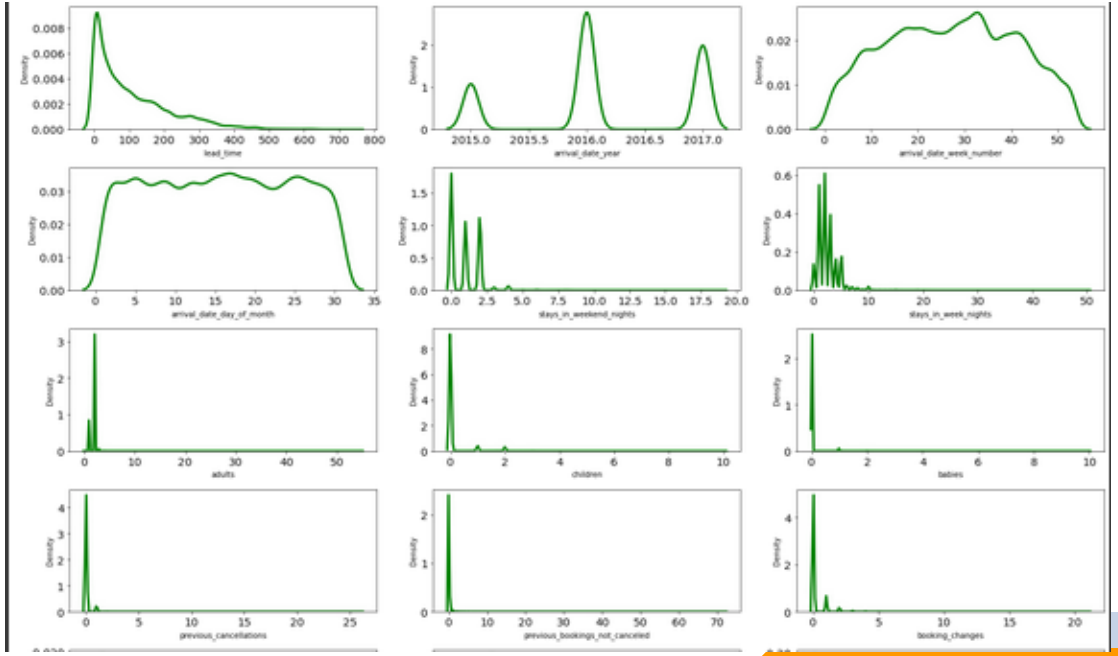
- Terlihat adanya outlier pada banyak kolom, yaitu lead_time, stays_in_weekend_nights,stays_in_week_nights, adults, children, babies, previous_cancellations, previous_bookings_not_canceled, booking_changes,days_in_waiting_list, adr, required_car_parking_spaces,total_of_special_requests.
- Pada boxplot terlihat ada terdapat banyak kolom yang kemungkinan skewed yaitu semua kolom kecuali is_canceled,arrival_date_week_number, arrival_date_day_of_month,is_repeated_guest



# UNIVARIATE ANALYSIS NUMERICAL (DISTRIBUTION PLOT)

## Insight:

- Seperti dugaan ketika melihat boxplot, hampir semua kemungkinan skewed kecuali is_canceled,arrival_date_week_number, arrival_date_day_of_month,is_repeated_guest. Berarti ada kemungkinan kita perlu melakukan sesuatu pada kolom-kolom tersebut nantinya.
- Ada 2 kolom yang terlihat mendekati normal yaitu arrival_date_week_number dan arrival_date_day_of_month.



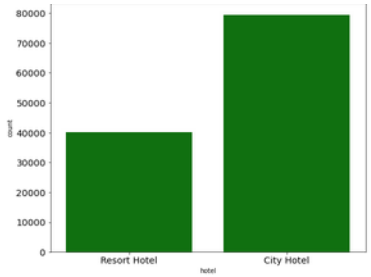
# INSIGHT UNIVARIATE ANALYSIS NUMERICAL

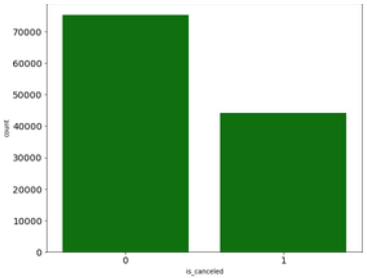
- Dari hasil terlihat bahwa tidak ada yang berdistribusi normal (distribusi normal : skew = 0).
- Hampir semua skew, baik positively maupun negatively skew, kecuali 'arrival date week number' dan 'arrival date day of month'.
- Ada 2 kolom yang mendekati distribusi normal yaitu 'arrival date week number' dan 'arrival date day of month'.
- Sehingga dugaan pada pengamatan boxplot dan kde plot terkait skew sudah terbukti.

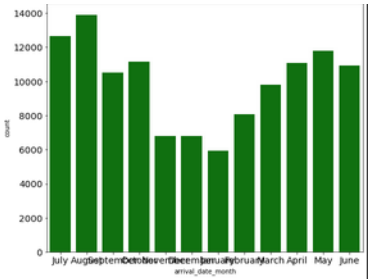
# UNIVARIATE ANALYSIS CATEGORICAL (INDIVIDUAL COUNT PLOT 1)

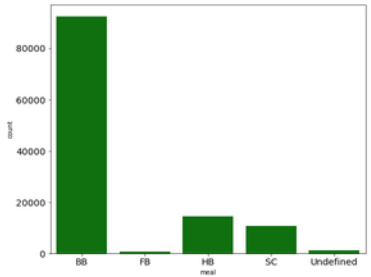
## Insight:

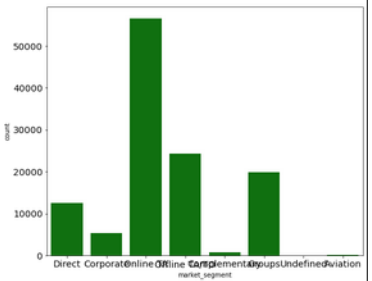
- Hasil dari pengamatan di atas di dapat kesimpulan antara lain hampir semua feature categorical terjadi ketimpangan yang cukup tinggi antara lain
- Hotel = Perbedaan city hotel dan resort hotel mencapai 50%
- Arrival date month = Pada bulan November - Januari mengalami penurunan yg signifikan dibanding bulan sebelumnya
- Meal = BB (Bed and Breakfast) sangat berbeda jauh dengan HB (Breakfast and Dinner, SC (Self Catering) dan FB (Breakfast, lunch and dinner)
- Market_Segment = Market Segment berbeda jauh dengan offline
- Distribution_Channel = Via TA/TO (Travel Agent, Tour Operator) sangat jauh perbedaannya dengan distribution channel lainnya
- reserved_room_type = type A sangat berbeda jauh dengan type lain (terdekat perbedaan lebih dari 4 kali
- assigned_room_type = sama seperti no 6 akan tetapi gap nya semakin kecil karena beberapa type mengalami peningkatan (G, E, D, F, B)



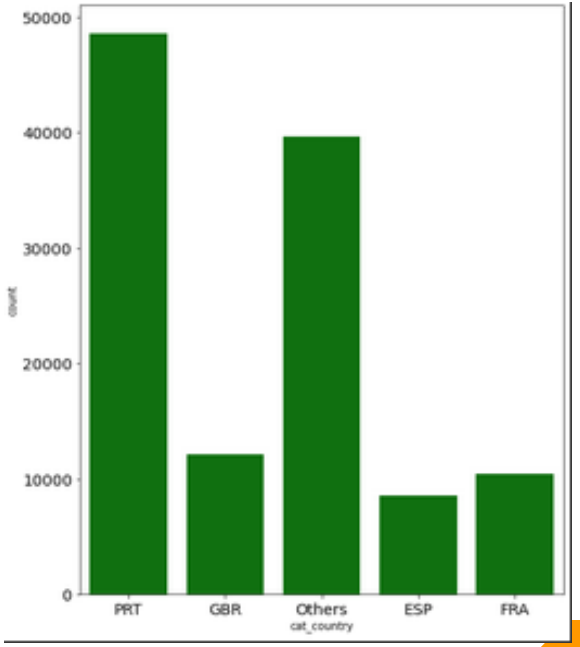




- Deposit type = No deposit masih tinggi penggunaannya dibanding dengan non refund atau refundable
- customer type = Transient dengan customer type lain berbeda sangat jauh

# UNIVARIATE ANALYSIS CATEGORICAL (INDIVIDUAL COUNT PLOT 2)

## Insight:

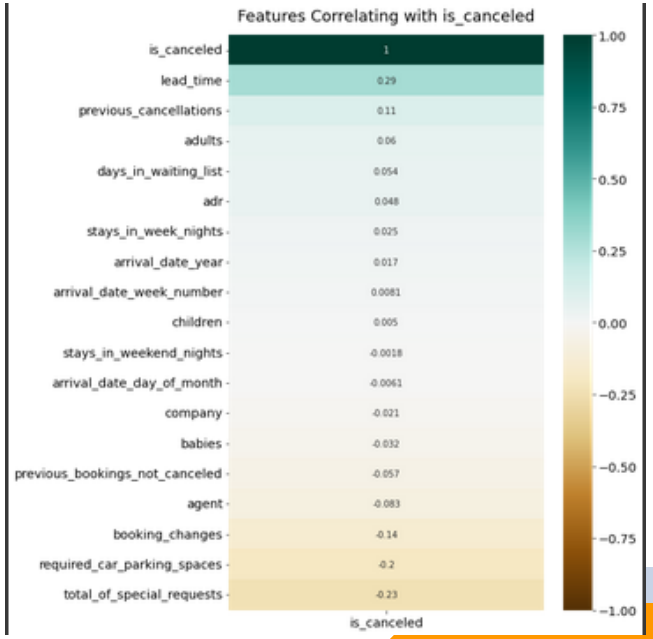
- Untuk features country juga memiliki nilai timpang dimana teringgi adalah turis dari PRT(Portugal) jauh lebih banyak dari country lainnya



# MULTIVARIATE ANALYSIS (HEATMAP CORRELATION)

## Insight:

- Target adalah is_cancelled yang menandakan status pembatalan pesanan
- is_cancelled sangat jelas memiliki korelasi negatif yang sangat kuat terhadap reservation_status dengan berat -0.92
- is_cancelled memiliki korelasi positif yang rendah terhadap lead_time, & country dirange 0,20 – 0,3999
- is_cancelled memiliki korelasi positif yang sedang terhadap deposit_type dengan berat 0,47
- Kolom market_segment & distribution_channel memiliki korelasi positif yang kuat sebesar 0.77. Kemungkinan redundan dan hanya menggunakan salah satunya
- Kolom agent & hotel memiliki korelasi positif yang kuat sebesar 0.79. Kemungkinan tidak digunakan karena tidak memiliki korelasi dengan target is_cancelled
- Kolom reserved_room_type and assigned_room_type juga memiliki korelasi positif yang sangat kuat dengan berat 0.81. Kemungkinan tidak digunakan karena tidak memiliki korelasi dengan target is_cancelled

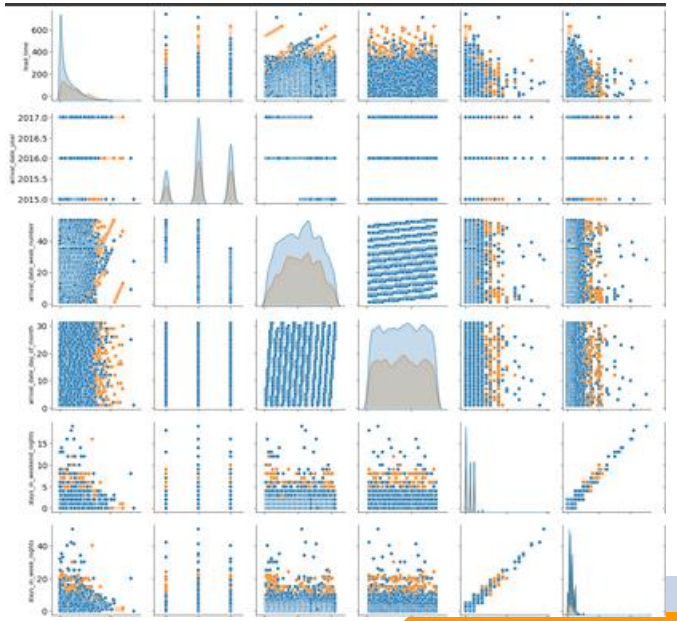


# MULTIVARIATE ANALYSIS (PAIRPLOT)

## Insight:

Setelah dibentuk pair plot dengan hue is_canceled, fitur yang menunjukkan pattern yang cukup jelas:

- stays_in_weekend_nights & stays_in_week_nights
- Terdapat korelasi positif antara stays_in_weekend_nights & stays_in_week_nights. Dimana Jika stays_in_week_nights tinggi, kemungkinan reservasi dibuat untuk durasi yang lebih lama (juga mencakup lebih banyak stays_in_weekend_nights)
- Selain dua fitur tersebut tidak terlihat cluster terpisah yang menandakan fitur saling berkorelasi.



# BUSINESS INSIGHT

EDA Conclusion

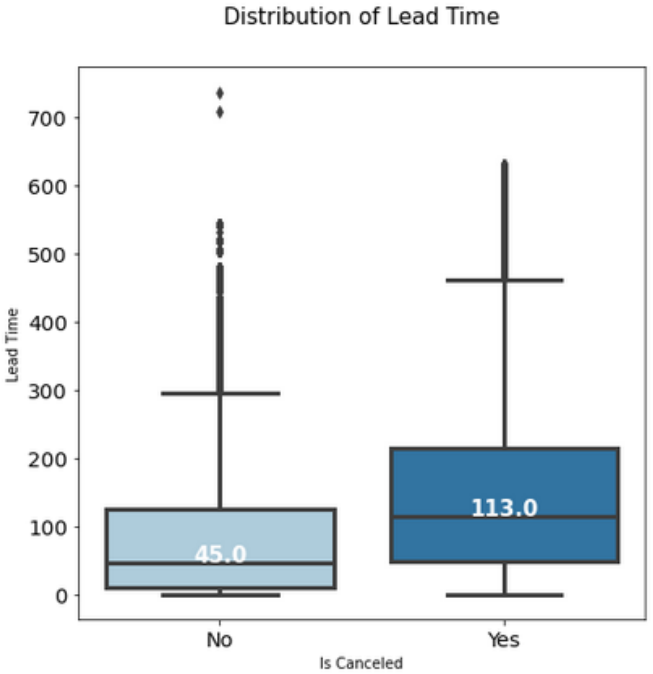
# LEAD TIME AND CANCELLATION

## Insight:

- Semakin lama lead time, semakin besar kemungkinan pelanggan akan membatalkan pemesanan hotel
- Jika jumlah hari antara saat pemesanan dibuat dan perkiraan tanggal kedatangan meningkat, kemungkinan pelanggan memiliki lebih banyak waktu untuk membatalkan reservasi dan lebih banyak waktu untuk keadaan tak terduga yang dapat membatalkan reservasi

## Business Recommendation:

- Hotel harus memberlakukan kebijakan untuk membatasi seberapa jauh hari pemesanan dapat dilakukan untuk mengurangi kemungkinan pembatalan



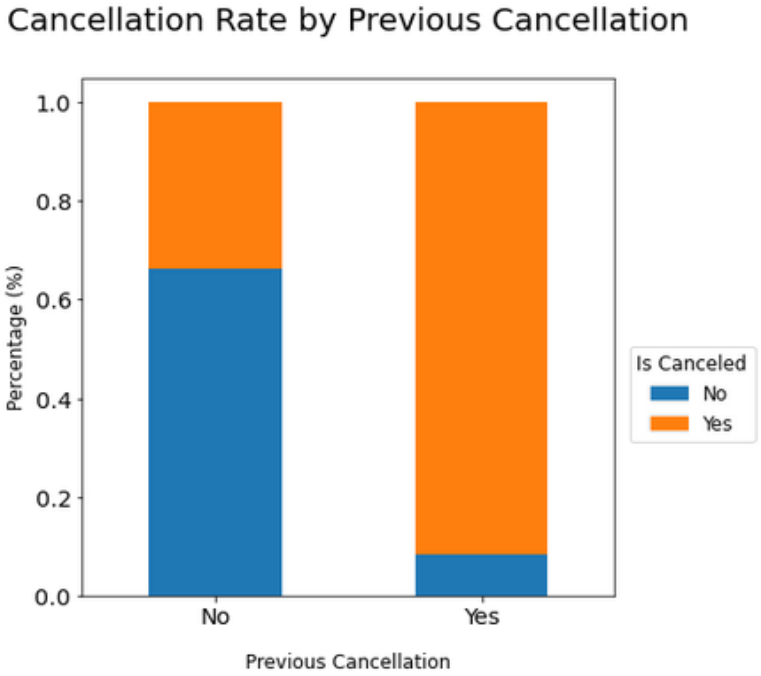
# DEPOSIT TYPE, PREVIOUS CANCELLATIONS OVER CANCELLATION

## Insight:

- Pelanggan yang telah membatalkan pemesanan hotel sebelumnya kemungkinan akan membatalkan lagi.

## Business Recommendation:

- Memberikan promo kepada customer yang sering melakukan cancel pada booking sebelumnya supaya customer tidak melakukan cancel



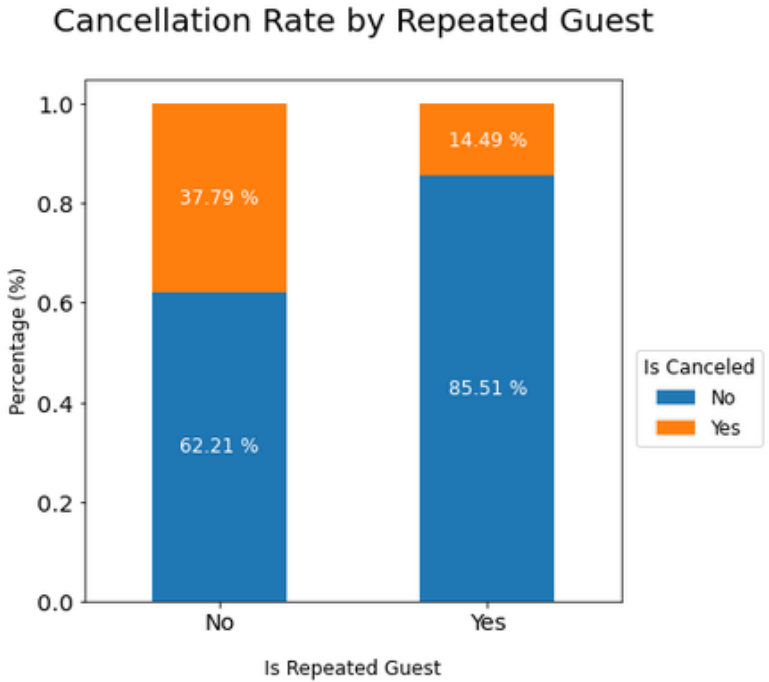
# REPEATED GUEST AND CANCELLATION

## Insight:

- "Repeated Guest" diidentifikasi memiliki peluang pembatalan yang lebih rendah.

## Business Recommendation:

- Hotel dapat menerapkan Kebijakan untuk memberi insentif (bonus/voucher,dll) & juga untuk meningkatkan lebih banyak "Repeated Guest" agar dapat kembali memesan kamar sehingga semakin bertambahnya tamu yang memesan ulang maka tingkat pematalan pesanan tunggal semakin minim



# TERIMA KASIH